Principles and Obstacles for Sharing Data from Environmental Health Research: Workshop Summary

DETAILS

114 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-37085-1 | DOI 10.17226/21703

AUTHORS

Robert Pool and Erin Rusch, Rapporteurs; Roundtable on Environmental Health Sciences, Research, and Medicine; Board on Population Health and Public Health Practice; Health and Medicine Division; National Academies of Sciences, Engineering, and Medicine

Principles and Obstacles for Sharing Data from Environmental Health Research

Workshop Summary

Robert Pool and Erin Rusch, *Rapporteurs*

Roundtable on Environmental Health Sciences, Research, and Medicine

Board on Population Health and Public Health Practice

Health and Medicine Division

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by contracts between the National Academy of Sciences and the Colgate-Palmolive Company, the ExxonMobil Foundation, The Kresge Foundation, the National Institute of Environmental Health Sciences, and Royal Dutch Shell. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-37085-1
International Standard Book Number-10: 0-309-37085-X
Digital Object Identifier: 10.17226/21703

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2016. *Principles and obstacles for sharing data from environmental health research: Workshop summary.* Washington, DC: The National Academies Press. doi: 10.17226/21703.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Ralph J. Cicerone is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.national-academies.org**.

PLANNING COMMITTEE FOR THE WORKSHOP ON PRINCIPLES AND OBSTACLES FOR SHARING DATA FROM ENVIRONMENTAL HEALTH RESEARCH[1]

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
KIMBERLY THIGPEN TART, National Institute of Environmental Health Sciences, Research Triangle Park, NC

[1] The National Academies of Sciences, Engineering, and Medicine's planning committees are solely responsible for organizing the workshop, identifying topics, and choosing speakers. The responsibility for the published workshop summary rests with the workshop rapporteurs and the institution.

ROUNDTABLE ON ENVIRONMENTAL HEALTH SCIENCES, RESEARCH, AND MEDICINE[1]

[1] The National Academies of Sciences, Engineering, and Medicine's forums and roundtables do not issue, review, or approve individual documents. The responsibility for the published workshop summary rests with the workshop rapporteurs and the institution.

SUSAN L. SANTOS, Rutgers School of Public Health, Piscataway, NJ
JOHN D. SPENGLER, Harvard School of Public Health, Boston, MA
G. DAVID TILMAN, University of Minnesota, St. Paul
PATRICIA VERDUIN, Colgate-Palmolive Company, Piscataway, NJ
NSEDU OBOT WITHERSPOON, Children's Environmental Health Network, Washington, DC
HAROLD ZENICK, U.S. Environmental Protection Agency, Research Triangle Park, NC

HMD Staff

KATHLEEN STRATTON, Study Director
ERIN RUSCH, Associate Program Officer (*until May 8, 2015*)
HOPE HARE, Administrative Assistant
ROSE MARIE MARTINEZ, Director, Board on Population Health and Public Health Practice

Reviewers

This workshop summary has been reviewed in draft form by persons chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published summary as sound as possible and to ensure that the summary meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the process. We wish to thank the following individuals for their review of this workshop summary:

Julia B. Brody, Silent Spring Institute
James W. Conrad, Jr., Conrad Law and Policy Counsel
Susan L. Santos, Rutgers University
Christine Thayer, National Toxicology Office of Health Assessment

Although the reviewers listed above have provided many constructive comments and suggestions, they did not see the final draft of the workshop summary before its release. The review of this summary was overseen by **Sue Curry,** University of Iowa. She was responsible for making certain that an independent examination of this workshop summary was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this summary rests entirely with the rapporteurs and the institution.

Contents

1

Introduction

On March 19, 2014, the Roundtable on Environmental Health Sciences, Research, and Medicine in conjunction with the Committee on Science, Technology, and the Law, both of the National Academies of Sciences, Engineering, and Medicine, held a workshop on the topic of the sharing of data from environmental health research. Experts in the field of environmental health agree that there are benefits to sharing research data, but questions remain regarding how to effectively make these data available. The sharing of data derived from human subjects—making them both transparent and accessible to others—raises a host of ethical, scientific, and process questions that are not always present in other areas of science, such as physics, geology, or chemistry. Through presentations from invited speakers and discussions among all attendees, the workshop participants explored key concerns, principles, and obstacles to the responsible sharing of data used in support of environmental health research and policy making while focusing on protecting the privacy of human subjects and addressing the concerns of the research community. The workshop statement of task is provided in Box 1-1.[1]

The following is a summary and synthesis of the presentations and discussions that took place during the workshop. When reading the summary, it is important for the reader to keep in mind that the opinions

[1] The planning committee's role was limited to planning the workshop, and the workshop summary has been prepared by the workshop rapporteurs as a factual summary of what occurred at the workshop. The statements, recommendations, and opinions expressed are those of individual presenters and participants, and are not necessarily endorsed or verified by the National Academies of Sciences, Engineering, and Medicine, and they should not be construed as reflecting any group consensus.

BOX 1-1
Statement of Task

An ad hoc committee will plan and conduct a public workshop featuring presentations and discussions to outline key concerns, principles, and prototype programs or best practices in fostering the responsible sharing of research used in environmental policy making while protecting the privacy of human subjects and addressing concerns of the research community. The committee will identify the specific topics to be addressed, develop the agenda, select and invite speakers and other participants, and moderate the discussions. A full-length workshop summary and a brief workshop summary will be prepared by a designated rapporteur in accordance with institutional policies and procedures.

expressed and any recommendations made are those of the individual speakers themselves and do not represent the position of the National Academies. Indeed, the purpose of the Roundtable on Environmental Health Sciences, Research, and Medicine is to provide a mechanism for interested parties in environmental health to meet and discuss sensitive and difficult environmental issues in a neutral setting. The Roundtable fosters dialogue about these issues, but it does not provide recommendations or try to find a consensus on these issues.

ORGANIZATION OF THE SUMMARY

The organization of this summary deviates somewhat from the structure of the workshop agenda. Specific comments made by speakers during some discussion sessions and some presentations have been regrouped to reflect overarching discussion topics, regardless of when they occurred throughout the day. When such deviations occur, they are noted. Chapter 2 introduces the terminology and the federal laws and policies that were presented in Session 1 of the workshop, as well as some presentations on current approaches to data sharing and the weaknesses of those approaches described throughout the workshop. Chapter 3 recaps the presentations that focused on the benefits of data sharing and the related discussions. Chapter 4 summarizes the presentations that explored the challenges associated with data sharing and the associated discussion. Chapter 5 primarily summarizes the discussions from Sessions 4 and 5 of the workshop, where participants offered insights into possible ways forward for the sharing of data from

environmental health research and overall reflections. The workshop agenda is found in Appendix A, and biographical sketches of the workshop speakers are included in Appendix B.

2

Current Approaches and Weaknesses of Those Approaches

This chapter is presented in two parts. The first part introduces some of the terminology used throughout the workshop, as well as the federal laws and policies that form the framework within which the sharing of data on environmental health occurs. After the summary of the presentations, relevant discussions from sessions from throughout the day[1] are described. The second part of the chapter summarizes the presentations that described current approaches to the sharing of environmental health data by federal agencies and identified the weaknesses and shortcomings of these approaches. Again, this is followed by a summary of the relevant discussion that occurred at the workshop.

TERMINOLOGY

Lynn Goldman, dean of the Milken Institute School of Public Health at George Washington University, introduced the topic of terminology in her workshop overview. "One thing that I have noticed is that some of the words that we use in science ... have sometimes not been used consistently," so she offered definitions for a series of terms that are important in talking about the sharing of environmental health data.

"*Peer review* is when you actually evaluate the scientific work by others in the same field," she said. "A *systematic review* is a summary of the clinical literature," she continued. "There are various methods that

[1] Presentations and, especially, discussion sessions often covered topics formally introduced in sessions different from the one being described in this summary. Where a discussion point from a different session is relevant, it is presented not in the order in which it was made but in proximity to the most appropriate discussion within this summary.

people can use. They can quantitatively pool the data or do a *meta-analysis*, which is a way of combining data from many different studies using a statistical process."

There are various approaches to testing and validating previous scientific work, she said. "A *reanalysis* is when you conduct a further analysis of data." A person doing a reanalysis of data may use the same programs and statistical methodologies that were originally used to analyze the data or may use alternative methodologies, but the point is to analyze exactly the same data and see if the same result emerges from the analysis.

"*Replication* means that you actually repeat a scientific experiment or a trial to obtain a consistent result," she continued. The second experiment uses exactly the same protocols and statistical programs but with data from a different population. The goal is to see if the same results hold with data from a different population.

"And then, finally, when you *reproduce*, you are producing something that is very similar to that research, but it is in a different medium or context," she said. In other words, a researcher who is reproducing an experiment addresses the same research question but from a different angle than the original researcher did. "Most of us, when we are doing systematic reviews, are more convinced that something is going on when we see reproducibility as well as replicability."

Different Meanings of "Data"

During Session 2, Bernard Lo, president and chief executive officer of The Greenwall Foundation, noted that there are a number of different types of data. There are raw data, which come straight from the survey or the experiment. There are cleaned-up data, which consist of the raw data modified to remove obvious errors. There are processed data, which are data that have been computed and analyzed to extract relevant information. There is the final clean data set that is provided with a publication. And there are the metadata that describe the data. All of these types of data are important in different ways and for different purposes (see Figure 2-1).

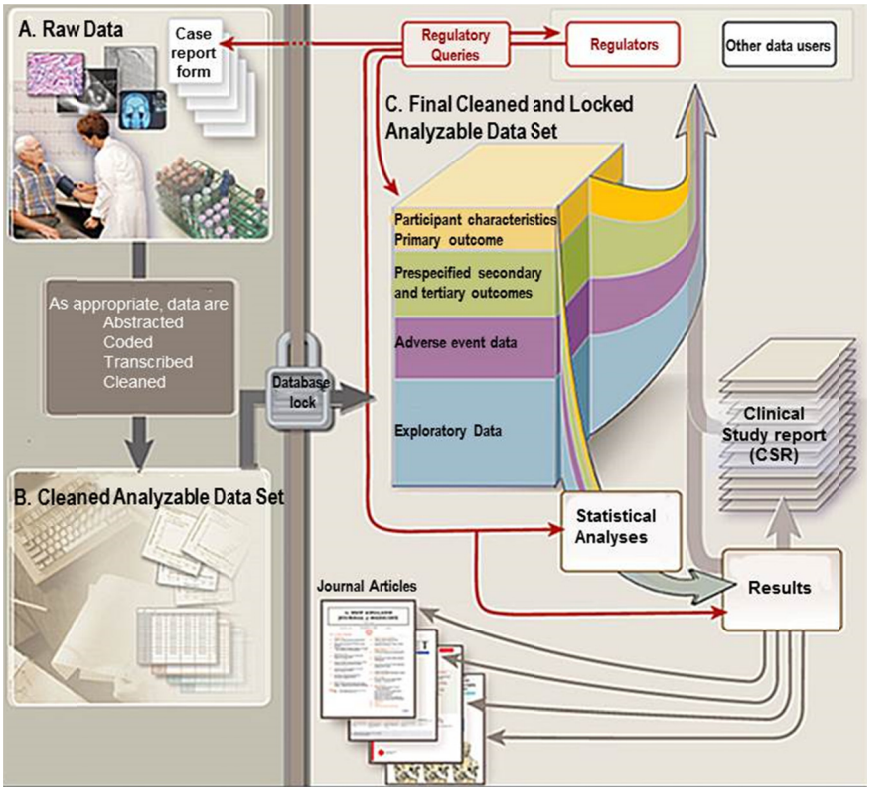

FIGURE 2-1 Data flow from participant to analyzed data and reporting.
SOURCE: IOM, 2014.

Lo explained that investigators may want to make the different types of data available to different people at different times. For each data-sharing element or activity, different questions may need to be posed. What types of data will be shared? Who should be providing data for sharing? Who should be the recipients of the shared data? When in the study process should certain data be shared and how? Should some people have open access? Is it important to pay attention to proportionality in terms of the risks, benefits, and burdens to various parties who are going to be affected by data sharing (e.g., researchers, participants, sponsors, the public)?

Lo noted that he chairs an Institute of Medicine committee that is tasked with issuing a report on the responsible sharing of clinical trial data. In January 2014 that committee issued a preliminary report, *Discussion Framework for Clinical Trial Data Sharing: Guiding*

Principles, Elements, and Activities, which lays out the committee's thoughts on the principles that should guide the sharing of clinical trial data, describes certain data-sharing activities, and defines the key elements of data and data-sharing activities (IOM, 2014). For its final report the committee's charge is "to analyze the benefits, challenges, and risk of various models of data sharing and to make recommendations to enhance the responsible sharing of clinical trial data," Lo said.[2] He added that while that committee is focused solely on clinical trials, many of its ideas and recommendations will likely apply to environmental health research more generally.

FEDERAL LAWS AND POLICIES PERTAINING TO DATA SHARING

A variety of federal laws specify what data must be shared and under what circumstances. In general, these laws apply to data held by federal agencies and to data collected with federal funding.

Paul Verkuil, chairman of the Administrative Conference of the United States, offered some background on these laws. "As a traditional matter," he said, "agencies were required to disclose data underlying investigations undertaken by agency scientists upon public request but were not required to disclose data from studies commissioned by the agency but performed by private entities. This framework has changed in recent years, and certain disclosure requirements also now apply to privately conducted research."

Specifically, Verkuil said, several federal acts govern the sharing of data. The Freedom of Information Act (FOIA)[3] requires that agencies release records—including scientific data—upon public request. It does contain a number of exceptions that protect things like confidential business information and personal privacy. The Electronic Freedom of Information Act Amendments of 1996[4] require agencies to release electronic copies of documents that have been previously requested and

[2] The Committee on Strategies for Responsible Sharing of Clinical Trial Data released its final report in January 2015. *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk* is available at www.nap.edu/catalog/18998.

[3] Freedom of Information Act, 5 U.S.C. § 552, Amended by Public Law 104-231, 110 Stat. 3048, 104th Congress.

[4] Electronic Freedom of Information Act Amendments of 1996, Public Law 104-231, 104th Congress.

that are likely to be the subject of future requests rather than waiting for subsequent requests. The result, Verkuil said, is much greater transparency than had previously been the case.

In 1998 a law referred to as the Shelby Amendment[5] was passed. That law required that all federally funded research data be made available to the public under FOIA. Traditionally, it was not the case that funded research had to be made available in response to a FOIA request. The Shelby Amendment changed that, requiring that data produced by grantees be released under FOIA, subject to the usual exceptions. The Shelby Amendment was enacted in response to concerns about a desire to reanalyze two studies, the Harvard Six Cities Study (Dockery et al., 1993) and the American Cancer Society Study (Pope et al., 1995) that were looking at the health risks caused by particulate matter in the air.

This was followed in 2001 by the enactment of the Information Quality Act, also called the Data Quality Act,[6] which was intended to improve the quality of information used and promulgated by agencies. Among other things, the act requires agencies to create a procedure that allows people to correct information that has been released if the information is erroneous, Verkuil said. Guidelines issued by the Office of Management and Budget (OMB) in response to the Information Quality Act state that when executive branch agencies provide "influential scientific, financial, or statistical information," they also "shall include a high degree of transparency about data and methods to facilitate the reproducibility of information by qualified third parties" (OMB, 2002). These OMB guidelines have affected how agencies respond both to requests for data and also to what are called "information corrections," which are intended to correct information that has been promulgated.

Verkuil noted that contractors and grantees are treated differently in the Shelby Amendment, with only grantees being forced to make data available. "I do not think that this distinction is one that can last long," he said, because once you get data from one type of federally funded research, it is difficult to imagine not requiring availability from all federally funded research. "The transparency is promoted when an agency relies upon a privately funded study. It urges the researcher to disclose the underlying data. When a private researcher declines, agencies should issue an explanation why they relied on such studies

[5] Shelby Amendment to the Omnibus Appropriations Act for Fiscal Year 1999, Public Law 105-277, 105th Congress.

[6] Data Quality Act, Section 515 of the Consolidated Appropriations Act for Fiscal Year 2001, Public Law 106-554, 106th Congress.

despite the declining. And agencies should require conflict-of-interest disclosures for all scientific research submitted to inform the decision-making process."

He noted that the ultimate beneficiary of such data sharing is the broader society. "Open communication among scientists and engineers and between these experts and the public accelerates scientific and technological advancement, strengthens the economy, educates the nation, and enhances democracy," he said.

The Role of Courts

In addition to Congress and the executive branch, including federal agencies, courts also play a role in determining how data are shared. Verkuil described that role in his presentation.

Most of the rules produced by the U.S. Environmental Protection Agency (EPA) and other federal agencies are developed through informal rule making, Verkuil noted, but when agencies issue rules through that process, reviewing courts scrutinize them very carefully. The process is governed by the "arbitrary and capricious" standard of 5 U.S.C. § 706(2)(A).

For an agency's rule to be upheld on judicial review, the agency must have placed in the administrative record all of the information that the agency relied upon to reach its decision—a requirement known as the "Portland Cement doctrine."[7] "The E-Government Act makes it a little easier," Verkuil said, "because you can post rules online and you can use regulations.gov to find out what else has been posted by other commenters." Sometimes, however, there can be problems with paper-based comments, Verkuil said. For example, they may not be scanned. This is a transitional problem, he said, but a real one for agencies depending on how many paper-based versus electronically transmitted comments that they receive. Verkuil commented that his organization, the Administrative Conference of the United States, encourages agencies to do as much as possible electronically rather than on paper.

This amount of information that can be presented to the court is huge. "You can appreciate what the 'record' looks like on review," Verkuil said. "It is enormous, and the agency has to decide what is in and what is out on its own." The situation is different from formal adjudication or formal rule making, in which there is an administrative

[7] The Portland Cement doctrine is named after the decision in *Portland Cement Association v. Ruckleshaus*, 486 F.2d 375, 393-94 (D.C. Cir. 1973).

law judge making the decision, the judge decides what goes in and what does not, and what is accepted becomes part of a "record" in the traditional legal sense. In contrast, Verkuil explained, in informal rule making, the record is an accumulation of the best estimate of what needs to be in there to support the rule.

Although courts can scrutinize the research and underlying data upon which the agency relied, Verkuil said, they generally defer to the agency on technical determinations. Courts recognize that they do not have the technical expertise to second-guess an agency's scientific or technical judgment. Instead, the courts seek to make sure that an agency has behaved rationally in light of the data before it. A court also looks to determine whether an agency has observed appropriate procedures in reviewing the underlying evidence and whether it considered relevant information and alternative approaches. Those procedures, Verkuil said, all seek to make sure that what an agency has produced was produced in an appropriate scientific manner, with proper judgments supported being by the evidence.

Finally, Verkuil touched on the issue of whether someone can take an agency to court if it is believed that the agency failed to comply with the Data Quality Act. He explained that this issue is still being debated and is yet to be resolved very convincingly. As you can imagine, he said, it could be a big issue if the courts decided to intervene every time that an agency decides whether the Data Quality Act has been properly complied with or not.

Executive Branch Guidance

More generally, the executive branch puts forward a variety of policies regarding data sharing. George Gray, director of the Center for Risk Science and Public Health at the Milken Institute School of Public Health at George Washington University, described some of the more relevant policies during his presentation. These policies are different from laws and regulations, he emphasized. They provide guidance and do not have the same force as either laws or regulations.

Many of the relevant policies are promulgated by the OMB, he said, and he focused on what are referred to as the OMB "circulars," which are instructions or information from the OMB to federal agencies that are generally in effect for 2 or more years.

The Shelby Amendment instructed the OMB to amend one of its circulars, Circular A-110.[8] In particular, the Shelby Amendment told the OMB to amend Circular A-110 to ensure that all data produced with funding from federal grants would be available to the public under FOIA.

In particular, Gray said, the amended version of Circular A-110 has the following provisions:

- It applies to grants and agreements with institutions of higher education, hospitals, and other nonprofit organizations.
- It obligates EPA to obtain from its contractors "research data" underlying findings used by the agency in developing action that has the force and effect of law.
- It has exceptions for drafts, peer reviews, personally identifiable information, and so on.
- It also exempts confidential business information or other information that needs to be confidential, but only until the data are published in a journal or cited by an agency in support of its action.

One of the things that is interesting about the circular, Gray said, is that rather than applying to research that is done within the federal government, it applies to grants and agreements that are made by the government with institutions of higher education, hospitals, and other nonprofits. In particular, it obligates EPA to get the research data that underlie findings that are used by the agency in developing agency actions. This is a way of building a record of what leads the agency to make a particular decision.

The OMB Circular A-130[9] lays out the basic principles that the Executive Office of the President wishes agencies to follow in using information to come to decisions. Those principles, as laid out in the circular, include the following:

- The free flow of information between the government and the public is essential to a democratic society. In other words, Gray paraphrased, "Sharing is the right thing to do."
- The nation can benefit from government information disseminated both by federal agencies and by diverse nonfederal parties,

[8] Circular A-110 can be found at http://www.whitehouse.gov/omb/circulars_a110 (accessed October 26, 2015).

[9] Circular A-130 can be found at http://www.whitehouse.gov/omb/circulars_a130_a130trans4 (accessed October 26, 2015).

including state and local government agencies, educational and other not-for-profit institutions, and for-profit organizations.

- The open and efficient exchange of scientific and technical government information, subject to applicable national security controls and the proprietary rights of others, fosters excellence in scientific research and effective use of federal research and development funds.

Finally, Gray described a memorandum from the Office of Science and Technology Policy to the heads of executive departments and agencies (Executive Office of the President, 2013). "Again, this is the center of the executive branch giving instruction, guidance, policy approaches to the other executive branch agencies," he said. The memorandum offered several of the administration's "policy principles," including the following:

- "The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community" (Executive Office of the President, 2013). "They want to see the direct results of federally funded scientific research made available to and useful for the public, industry, and the scientific community," Gray commented. "Again, this is the exhortation to more data sharing, more openness in the way things are done."
- "Scientific research supported by the federal government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security" (Executive Office of the President, 2013).

"These are the policies," Gray concluded. "This is, again, the center of the executive branch speaking to all of the executive branch agencies, the ones that ultimately end up implementing all the various laws that come out of Congress."

COMMENTS AND DISCUSSION

Several challenges related to the current approaches to data sharing were highlighted during the Session 1 discussion. Workshop speakers and participants provided individual remarks that are summarized in this section.

The Limits to Data-Sharing Requirements

Given that the federal government requires researchers to share data that have been collected through the use of federal funds, several workshop participants raised the issue of just how far that requirement extends. Does it extend to any research project that has accepted any federal funds for any aspect of the project? Does it extend to research that was done 10 or 20 years ago?

Gwen Collman, the director of the Division of Extramural Research and Training at National Institute of Environmental Health Sciences (NIEHS), noted that the National Institutes of Health (NIH) has data-sharing policies for the researchers that it funds. "We require data-sharing plans for some of our investigators depending on the size and scope of the funding that they receive," she said.

Goldman pointed out that "[t]here are many statutes now that require regulatory agencies to use [the] best available data," she said. "The agencies are not supposed to simply use data that are submitted to them, but they are supposed to do a data dragnet and find the best available data whether these are data that the investigators want to submit or not." But this does not take into account the researchers' wishes, she said. "I think a concern by investigators has been, 'I did not do that research for the purpose of a regulation. Now I am being asked to undertake the burden of doing all these special things for a regulatory agency that did not fund me.'"

The issue becomes even more complicated if the best available research was done by researchers in other countries. So, Goldman asked, "What is the obligation of investigators to a regulatory agency that perhaps did not fund them?" She offered as an example an investigator in Norway who did a study that EPA decided was one of the best studies on a particular contaminant, so this study, as long as it was published in the scientific literature, is supposed to be included in the EPA assessment.

"I do not think you can subpoena the investigator in Norway and get his data," Verkuil said, "but it does raise a bit of a conundrum for the agency." That conundrum centers on the precise meaning of the word

"consider," because anything that an agency considers is supposed to be put in the record on review. "Now what is 'considering'?" he asked. "If you are doing a proactive review of the scientific literature before you make a decision and you see something in a Norwegian scientist's report, are you 'considering' it? ... If the scientist does not have, let's say, underlying codes available and other things, which means it cannot be analyzed, then it probably would not have been considered. But I do not know how you go beyond that."

Linda Birnbaum, director of NIEHS of NIH, offered as an example a project in which NIEHS funded a small piece of an analysis of some data that had been collected by Norwegian investigators. "It is often unclear to us, even given the Shelby Amendment, whether that data has to be provided upon FOIA or not," she said. If the analysis is being used in rule making, one interpretation would be that because at least part of the analysis was federally funded, then the data would need to be provided, but she added, "It is certainly not crystal clear to us what our grantees have to make available in those situations."

A second type of complication revolves around the issue of who owns the data being requested. As Steven Lamm of Consultants in Epidemiology & Occupational Health, LLC, noted, "So many of the papers that we use were somebody's Ph.D. dissertation, and they are now in some other institution, or they are not interested in that area anymore, and the professor has moved on to other things. In that case, he said, it can be exceptionally difficult to, first, find out where the data are and, second, discover who is in a position to release the data. Those are critical field-level issues," he said.

Alan Morrison of the George Washington University Law School agreed with Lamm and added that the issue becomes particularly tricky for research done at state universities when the questions arise as to whether the researcher owns the data or the university owns the data and what happens when there is a conflict.

Lamm added that a related issue concerns reimbursement for supplying the data. "If the agency is going to request data, it ought to have a budget that allows it to pay for the acquisition," he said. "Projects have been funded. The budgets no longer exist. Asking somebody to go and find the data in the archives requires time and money."

Dan Greenbaum, president of the Health Effects Institute, explored the issue further. "Going back to the study of Norway: suppose it is one of a dozen studies that have found similar things, maybe some positive, some negative, but the agency is considering it as one of several. Or

suppose that study in Norway is one of two studies worldwide that have found an effect that the agency is trying to characterize and think about regulating. Is there a different legal standard there?" How would an agency approach those two different situations? And how would a judge think about them?

Verkuil suggested that a judge would likely rule differently on the two situations. "If it is one of two studies, then you have to have it, and, as the agency, you better well track it down and have it ready. If it is one of ten, you explain why you did not do it, but you do not need it, I think." In short, he said, you use logic to determine how important a study is to a particular decision.

Morrison added that it would be important whether there were any countervailing studies. That is, the role that a study plays in an agency's determination vis-à-vis other studies is an important factor in determining whether a study must be included. "It is very context specific," he said. "You really need to go out and try to get the data," but if you cannot get the data that you want, you get the data that you can. "Many statutes require you to use the best scientific evidence available," he said. "You may not be able to get the very best," so you get the best available.

Goldman suggested that there are other considerations to take into account as well. "What is the seriousness of the outcome? Is this something that kills people or gives them a slight headache? Also, where was it published? The best journal in the world?" It is necessary to take into account the quality of the peer review and the judgment of scientists about the quality of the study.

Finally, it is important to think about when the study was done. "There are many things that have been demonstrated in the environmental literature decades ago that you could not get published today," Goldman said. "Benzene and leukemia—nobody is seriously able to even study some of those. Asbestos and lung cancer—the data for those are not available in raw form. If they are available anywhere at all, they are probably in media that you cannot read anymore. You cannot utilize them. To say I am going to lop things off and only allow the use of data that you can actually acquire is difficult to hear in that context."

What Exactly Is Required When Sharing Data?

Morrison raised the question of exactly what is expected to be provided under data-sharing requirements. "In addition to data that are

actually produced, there are a lot more things, stuff that is out there as part of the process," he noted, "for example, the original proposal to a federal agency to do a study. There is the protocol that was developed. There are algorithms. There are models that were used by a grantee. Do those have to be made available ... to agencies, and should they be made available?" After all, he commented, understanding the data requires more than just having access to the data themselves; all of these other pieces may play a role as well.

Gray answered that scientific journals generally require authors to provide whatever things were used to produce the results, such as "the raw data. Many journals will require you to post your computer code, whether it is an analytic code in SAS [Statistical Analysis System] or you write your own code to help a particular model to get a result." Gray said he was not certain whether the Shelby Amendment would require computer codes to be provided by researchers who received federal funding.

Goldman offered her own thoughts on the issue. "Speaking as an epidemiologist," she said, "most of us believe that our questionnaires and protocols are fair game and that we do need to make them available when people want to see them. There are times when if you do not see the questionnaire and understand exactly how the data are collected that you really cannot understand what the responses mean. It is the art of epidemiology. Depending on how you ask the question [and whom you are asking], you can get a very different answer."

FEDERAL AGENCIES

It is, then, up to the various federal agencies to carry out the laws passed by Congress and to follow the instructions provided by the OMB and other executive agencies. During different sessions throughout the day, four presenters offered details about what specific federal agencies do in response to these laws and directives.

U.S. Environmental Protection Agency

Goldman spoke about EPA's experience with data sharing under the Information Quality Act (IQA) and Circular A-110 during her opening remarks. "A couple of years ago, a couple of us looked at the experience at the Environmental Protection Agency under the IQA and found that according to EPA's Web page at that time, over 10 years there were 79

requests that had been filed," she said. "Of these, only two actually asked for raw data. The request for raw data was not a common request."

Goldman then offered some details about the two requests. The first was a case that involved a perchlorate study, and the requester was an industry consortium called the Perchlorate Study Group. EPA had the relevant data from a contractor, and it made the data available to the study group, which "allowed them to examine the original brain images from the animals that were studied in this study as well as the original contractor's reports that actually contained data tables."

The second case was not so straightforward. In 2008 an industry group called the Association of Battery Manufacturers asked for raw data from a systematic review of a number of studies of lead toxicity. The principal investigator, Bruce Lanphear, had solicited colleagues all over the world to provide raw data from studies that they had carried out on the effects of lead toxicity on the intelligence quotient of children (this example is also referenced in the Chapter 3 discussion). EPA provided Lanphear with some of the funding for his systematic review, and the agency then used the result of this review in developing a lead-in-air standard.

"I spoke with Bruce about this [request for raw data] at the time we did our review," Goldman said. "He had signed data transfer agreements with these investigators promising that he would not release these data to other people.... He felt that those agreements precluded him from sharing those data with anyone else. However, EPA ruled that the data needed to be made available pursuant to the Shelby Amendment because of the fact that some federal funding had been made available for doing this systematic review." There were a few other factors as well, Goldman said, such as the fact that the battery manufacturers were suing EPA about the National Ambient Air Quality Standards, and EPA did not want to act until the lawsuit was settled. "But at the end of the day," she said, "EPA prevailed upon Cincinnati Children's Medical Center, and Bruce's hard drive was taken and provided to EPA."

Currently, Goldman continued, there is a new request for EPA to release raw data from the Harvard Six Cities Study and the American Cancer Society Study. It is from Senator David Vitter, who provided a statement on the case on March 11, 2014:

> As the input and output files are fundamental to conducting reanalysis, I repeatedly requested that EPA (1) obtain all the data files; (2) determine which data files pose a threat to privacy; (3) immediately release all

> data files that do not pose a threat to privacy; and (4) investigate measures to remove all personal health information from the files that contain confidential data prior to release. (Vitter, 2014)

"I think we are going to hear a lot more about this situation," Goldman said.

Furthermore, in February 2014 a bill[10] was introduced in the House of Representatives to "prohibit the U.S. Environmental Protection Agency from proposing, finalizing, or disseminating regulations or assessments based upon science that is not transparent or reproducible." Specifically, Goldman said, the bill is aimed at making sure that EPA specifically identifies all scientific and technical information used in proposing, finalizing, or disseminating any action and that it makes such information "publicly available in a manner that is sufficient for independent analysis and substantial reproduction of research results." The actions covered by the bill, Goldman said, are not just regulations but any risk, exposure, or hazard assessment; criteria document; standard; limitation; regulatory impact analysis; or guidance—in effect, almost anything that the agency does.

National Institute for Occupational Safety and Health

During Session 3 of the workshop, John Howard, director of the National Institute for Occupational Safety and Health (NIOSH), offered six principles for the sharing of data based on lessons learned at NIOSH. Overall, he said, "Scientific data developed with taxpayer dollars by taxpayer-supported scientists should be shared with data requestors unless there is a strong countervailing interest that can be articulated [and] that supports a decision to withhold data in a manner that prevents data reanalysis." In short, the sharing of data should be the default position, and if it is not feasible to share all of the data—for instance, because of privacy concerns—then one should share as many of the data as possible.

His first lesson is that "researchers need to think about the optimal data-sharing practices at the study concept stage." There should be a balance between the ability of the investigators to complete the research mission and the ability of legitimate data seekers to have timely access to data from the study that can be used for reanalysis. "Frequent communication with

[10] H.R. 4012, Secret Science Reform Act of 2014, 113th Congress.

data seekers during the study, while it is happening, is really highly recommended," he said.

Lesson two is that research budgets should prospectively include sufficient resources for the investigators to implement robust data-sharing plans. "NIOSH has found that implementation of data-sharing plans can be resource intensive," he said.

Lesson three is that data use agreements have limited value. "While they are somewhat popular," he said, "they do not provide particularly strong protections for sensitive, potentially identifying data provided to data analyzers because it is difficult to monitor and it is difficult to enforce specified restrictions on data use. Recognizing these limitations, we found that data use agreements need to clearly state what happens in the case of nonadherence."

Lesson four is that secure enclaves can protect the confidentiality of highly sensitive and potentially identifying data sets while providing access for analysis that is maximally useful. He noted that "the data can be used within those enclaves, but only nonidentifying aggregate analysis can be removed from the enclave."

Lesson five is to ensure that reanalysis is based on a strong and reproducible foundation. "Only clean, verified, finalized data sets from completed and published studies should be shared," he said. "We found that sharing preliminary data sets, which were not the basis of the published study, definitely confuses the data reanalyzers and creates scientific confusion in the end."

Lesson six is to make certain that study participants are not surprised by data disclosure issues that can arise from reanalysis. He noted that study participants should be made aware of the possible scope of disclosure parameters at the time of enrollment in the study that they are actually going to participate in.

National Center for Health Statistics

Edward Sondik, former director of the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention, spoke about data sharing at NCHS during Session 4 of the workshop. "Our role is to provide information for policy and research—information about the health care system and about the health of people in this country as compared with other countries," he said. Section 306 of the Public Health Service Act describes the general NCHS mandate by saying that the center "shall conduct and support statistical and epidemiological

activities for the purpose of improving the effectiveness, efficiency, and quality of health services in the United States."

A second part of this mandate, Sondik said, is to ensure the widespread dissemination of and access to the data that it collects. NCHS collects a wide variety of data. It collects health and health care data in such surveys as the National Health Interview Study and the National Health and Nutrition Examination Survey, it coordinates and collates data concerning births and deaths in the United States, and so on. "All of that is extremely important," Sondik said, "but if we put it in a safe, it does absolutely no good at all." Thus, the center's prime directive includes not only the collection of data but also the dissemination of the information that it collects.

"Then we have another prime directive, which is about confidentiality," he said. "It is really clear. It prohibits the release of potentially identifiable data. This is under Section 308(d) of the Public Health Service Act. Other agencies have their own confidentiality legislation. But in general, over the last several years it has been covered by the act we called CIPSEA, the Confidential Information Protection and Statistical Efficiency Act."

There are serious penalties for violating confidentiality, he said—fines of up to $250,000 or up to 5 years in prison. "I took this very personally. This is what it says under 308(d): No information identifying the person supplying the information may be released in any form without the consent of the person. It is really clear."

NCHS disseminates information in a variety of ways, including publications and public-use data files. Essentially everything that the center provides is put on the Web now, Sondik said. In addition to the publications and data files, there are a number of data access tools and also various linkages between the data, such as links from one survey to another. "We have some very interesting things that we do along that line," he said.

NCHS maintains a balance between the widespread dissemination of the data and maintaining the confidentiality of the people in the data files. The center uses a number of strategies to maintain confidentiality, Sondik said. "First of all, we create public use data sets which are as deidentified as we can make them.... We try to do the best job we can and produce public use data sets that anybody can use [and] that will safeguard the identity of the people who supply the data." However, he noted, there is no way to know exactly what the probability of disclosure is for any of the information in the data sets. "I wanted a probability of

disclosure for 17 years, but we do not have probability of disclosures. We do not know what those probabilities are."

To allow researchers to work with sensitive, personal, identifiable data, NCHS uses a variety of approaches, said Sondik. It developed research data centers that researchers can come to to work with the data or else can access remotely. It has also used other data enclaves, and it has reworked the data sets in ways that change the data enough to protect the identity of the individual respondents but not so much that the data are no longer useful to researchers.

A strategy that NCHS has chosen not to use is licensing, which provides identifiable data to researchers under a licensing agreement that prohibits them from sharing the data publicly or sharing it with unauthorized users. The National Center for Education Statistics does use this approach, Sondik noted. "That is something that we at NCHS felt that we were not able to do because of our ... interpretation of the legislation," he said. "It is interesting. You have two federal statistical agencies taking different strategies."

Given the potential effects of reidentification on individuals and the possibility that such risks could keep people from taking part in environmental health studies, Sondik suggested that the effects of disclosure should also be discussed in terms of the ability to carry out research in this area. "There is impact on the individual," he said, "but from my viewpoint, it was impact on the agency and our ability to continue to function and to be able to collect very sensitive information and preserve that information" that merits consideration.

Sondik discussed a particular issue that must be taken into account when an attempt is made to understand the likelihood of reidentification of the data in a data set. "The mosaic problem," he explained, "is the fact that there is this semi-infinite set of data sets out there, and data in one can relate to another, which can relate to another, which can relate to another. You can start with some piece of information, relate that to a different data set [and] to another data set and wind up, after you go through this sequence, being able to actually identify a person supplying data in that first data set." Because of the complexity of the issue and the uncertainties surrounding exactly what data are available, he said, it is extremely difficult to get any sort of estimate of the risk of reidentification through such an approach.

Finally, Sondik suggested that in understanding the likely response of the public to the risks of reidentification, it is important to understand what the public expects in terms of privacy, particularly in this era when

people are willingly offering up more and more personal information in various venues. "There has been some work on the part of the federal statistical agencies to understand what the public expectations are," he said, "but I do not think we have a very good handle on it." In particular, no one has explored the "inconsistency" between the information that people allow various entities—Facebook and other websites, for example—to collect and to use and the thought that these same people expect absolute confidentiality when it comes to the information that they provide to scientific researchers.

National Institutes of Health

Birnbaum described several NIH efforts focused on data during her presentation in Session 5 of the workshop.

For one, NIH is taking the lead on working to improve the reproducibility of data, Birnbaum said. In February 2014, NIH director Francis Collins and principal deputy director Lawrence Tabak published a commentary in *Nature* (Collins and Tabak, 2014) that discussed the lack of reproducibility in health research, especially preclinical studies, and described what NIH is intending to do to address the issue. For instance, NIH is developing a mandatory training module on the responsible conduct of research that will be given to NIH-funded trainees, both intramural and extramural, and the various institutes and centers at NIH are developing checklists to ensure the more systematic evaluation of grant applications. Pilot programs are also being run to assess the value of such things as cross-reviewing panels. "You take one reviewer from each panel," Birnbaum said, "and have those reviewers look at the whole review on another ongoing panel, and they also have the specific task of evaluating the scientific premise of the application. In other words, when they are reviewing a grant, [they ask concerning] the key publications on which an application is based, Are these in fact valid or appropriate publications?"

NIH also has a major initiative called BD2K, an abbreviation for Big Data to Knowledge. The goal of this $60 million project, Birnbaum said, is to make biomedical data intelligible, accessible, and citable. As part of the project, four centers of excellence are being established for biomedical big data analysis.

Birnbaum also described efforts at her own institute, NIEHS, aimed at sharing environmental data. "Our journal, *Environmental Health Perspectives*, was a pioneer in open access for scientific journals, which

is now being called for by legislation in the House," she said. "Our policy for open access is decades old, and it predates the policies from *PLoS* (the *Public Library of Science*), NIH, and PubMed requirements, for example. And we are evolving publication policies to deal with data-sharing issues, such as all the supplemental materials that are online." The journal is also working to improve reproducibility by, for example, requiring a checklist to be filled out when a paper is submitted to the journal to make sure that some of the key information related to study design is clearly presented in the study.

COMMENTS AND DISCUSSION

This section is a summary of the discussions related to federal agency implementations that took place throughout the workshop.

Examples of Secure Enclaves

During the discussion after Session 3, Howard further described the secure data enclaves with which NIOSH is experimenting. The work is part of an effort to make data that have been used in analyses accessible to people who wish to look at the data themselves to verify that an analysis was accurate. "We are trying to figure out how to make [those data] accessible while maintaining ... privacy, not just personal privacy, but also trade secret issues and several other avenues of privacy," he said. The data enclaves are one avenue that NIOSH is examining in order to make data available in this way, Howard said, but the institute's work with data enclaves is not far enough along that he could report how well they work. "It is fairly new," he said. "We are probably not at the stage where I could report that it is entirely worked out. That remains to be seen."

In contrast, Greenbaum noted during Session 4 of the workshop that other federal agencies do have working data enclaves. "The research data centers, which NCHS runs in Hyattsville, Maryland, are set up," he said. "Investigators do go in. They get access to NHANES [National Health and Nutrition Examination Survey] data and other information that [aren't] just generally available." These data centers are secure facilities where researchers carry out analyses on the data that are kept there. The centers provide various statistical packages for the researchers to use. "You cannot take the data set back out," he said, "but there is a mechanism for doing it in the federal government. How well that will

work with every single study that the federal government has funded is where we are still in the pilot stages of figuring out."

Looking to Scientific Journals

During Session 1, Gray highlighted the fact that *PLoS*, the first and the biggest online journal, has just published new data policy procedures that focus on public access to the data. "Access to research results, immediately and without restriction, has always been at the heart of *PLoS*'s mission and the wider open-access movement," he said. "However, without similar access to the data underlying the findings, the article can be of limited use. *PLoS* is trying to increase access to data and is revising its data-sharing policy." Authors who submit to *PLoS* must now make all data publicly available without restriction immediately upon publication of the article, he said. "Going forward, I think we are going to see a different world. This is being driven by a lot of forces."

Furthermore, as Birnbaum noted in Session 5, the NIEHS publication *Environmental Health Perspectives* was one of the first journals to require the authors of manuscripts to make the data in their papers available in a database, but many journals have followed suit, and that is becoming standard practice in the scientific publishing industry.

Francesca Dominici, professor of biostatistics and senior associate dean for research at the Harvard University School of Public Health, said during the discussion after Session 2 that there are journals in biostatistics that require authors to make both their data and their software available to others. On a more global level, Gray had said in Session 1 that the scientific world is changing rapidly, driven by the increasing ability to move information around electronically and a growing desire for openness in the scientific community. "If you want to publish in *Nature*, one of the very best journals in the world, [you] are required to make materials, data, and associated protocols promptly available to readers without undue qualifications. They want to see data sharing there."

Information Concerning Data Collection and Analysis

In Session 2, Collman noted that many of the difficulties in reproducing studies are related to the methods used and simply sharing data will do little to help. The scientific papers that describe a study often have relatively short methods sections that are not very detailed, and

someone in another laboratory who is trying to replicate those studies from scratch can find it quite difficult.

When there are human data from epidemiological or clinical studies, Collman said, the real question is: What kinds of communication are necessary to fully communicate the details and the nuances of how those studies were done as well as what all of the data mean and how the data variables are created? "We do have metadata, and we have data dictionaries," she said. "But oftentimes ... the methods of how we recruit and how we select participants and what the demographics of the original group are is not the most exciting paper to write. But in thinking about a future world where those data that come from those things end up in an open-access database with the proper protections or are available for sharing with consent of the research group, ... we really need to pay a lot of attention to the details of how these things came to be in order for the next group of scientists to be able to use them."

In addition to getting access to the data and to information about how the data were collected, it is also important to have access to information about how the data were processed and used to come to a decision, said Gray in his presentation during Session 1 of the workshop. He acknowledged that one of the factors that causes people to question how a governmental agency came to a decision is the unavailability of the data that underlay the decision. If data are unavailable—for example, because they are considered to be confidential business information—then some people may be concerned that "the agency is playing around with data when they are doing their analyses" and that "some part of data that might be important is not being released."

However, Gray continued, "I would say the place where I had the greatest trouble with this is actually understanding the agency's underlying reasoning. How did the data that were available to the agency end up in this result?" In using data and scientific information to come to conclusions, there are always many choices to be made, Gray said. "What models are we going to use? Which populations are we going to use? Which studies will we rely upon? Which ones will we not rely upon? All of those can ultimately have an influence on the end product, especially if the end product is something quantitative, like a national ambient air quality standard or a reference dose in EPA's integrated risk information system." Thus, it is crucial when sharing data to also share how those data were used in coming to a particular conclusion.

REFERENCES

Collins, F. S., and L. A. Tabak. 2014. NIH plans to enhance reproducibility. *Nature* 505(7485):612–613.

Dockery, D. W., C. A. Pope, X. Xu, J. D. Spengler, J. H. Ware, M. E. Fay, B. G. Ferris, and F. E. Speizer. 1993. An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine* 329:1753–1759.

Executive Office of the President. 2013. Memorandum for the heads of executive departments and agencies. Available at http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo__2013.pdf (accessed October 26, 2015).

IOM (Institute of Medicine). 2014. *Discussion framework for clinical trial data sharing: Guiding principles, elements, and activities*. Washington, DC: The National Academies Press.

OMB (Office of Management and Budget). 2002. Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by federal agencies. Available at http://www.whitehouse.gov/omb/fedreg_reproducible (accessed October 26, 2015).

Pope, C. A., M. J. Thun, M. M. Namboodiri, D. W. Dockery, J. S. Evans, F. E. Speizer, and C. W. Heath. 1995. Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *American Journal of Respiratory and Critical Care Medicine* 151:669–674.

Vitter, D. 2014. Memo to Dr. Francesca Grifo. Available at http://www.epw.senate.gov/public/_cache/files/6de2a2b9-ad38-41bc-a0c4-c909b391a526/031714 vitterlettertodrgrifodatamisconduct.pdf (accessed October 28, 2015).

3

The Benefits of Data Sharing

This chapter describes presentations at the workshop that focused on why data sharing is important, especially in the frame of the current state of environmental health sciences research. The relevant discussion from the workshop is summarized in the final section of this chapter.

ADVANCES FROM REPRODUCIBLE RESEARCH

Francesca Dominici, professor of biostatistics and senior associate dean for research at the Harvard University School of Public Health, discussed the importance of advancing science through reproducible research. She explained that a spectrum exists with data sharing: publication is at the low end of the spectrum, the codes and data underlying the published work are in the middle, and full replication of a study is at the highest part of the spectrum. "I think it is extremely hard to achieve a full replication in environmental health research for all kinds of reasons." For example, she said, there is confusion around what is meant by data. Additionally, raw data may be confidential, or some investigators have their entire careers built on the raw data. However, reproducible research is still achievable, and much progress in this area can be made by focusing on the most important data and the most important results.

Dominici made a distinction between reproducibility and replication, noting that "replication is not always strictly necessary" and often requires tremendous investments. One aim of the reproducibility standard is to fill the gap in the scientific evidence-generating process between full replication of a study and no replication. Utilizing the tools available, investigators can reproduce the results (and verify the quality of scientific claims) when some data and analysis codes from the original

research are made available. A researcher does not need access to the raw data because statistical analyses can be performed on the aggregate data and the methods and results can be shared.

Reproducible research still requires funding and energy to maintain access to the data and software, and this burden often falls on the original investigators. Dominici highlighted the need for an established system to promote reproducible research that could be sustained with support from government agencies.

ALLOWING VERIFICATION OF RESULTS

One of the major factors behind the push to share environmental health data is the widespread use of such data to develop environmental regulations. For environmental policy making to be legitimate, the scientific reasoning behind a given decision—including the data supporting it—must be transparent. Debates over policy can become quite adversarial, and that adversarial nature can extend to debates over the data and their interpretation.

There is a generalized concern over how trustworthy scientific findings are, particularly in controversial fields, as noted by Anthony Cox, chief science officer of NextHealth Technologies. "I think that this issue of whether we should share data and how we should share [them], two different issues, should be framed in the context of where we are right now," he said, "and where we are right now is that there is an epidemic of false-positive and irreproducible results in the peer-review literature." To illustrate, he referenced three different articles focused on perceived problems in science today and the errors that are often found in published results that caused much hand-wringing:

- "Why Most Published Research Findings Are False" (Ioannidis, 2005),
- "Trials and Errors: Why Science Is Failing Us" (Lehrer, 2011), and
- "Beware the Creeping Cracks of Bias" (Sarewitz, 2012).

There are many reasons why different investigators might reach different conclusions from the same data, Cox said, so it is important not only to share data but also to share the reasoning that led to a particular conclusion to assess if any mistakes were made. For example, he said, "I would say that if a regulatory agency says we are going to do this because

the relevant risk is 2.2, the relevant data are the complete audit trail that ends with 2.2 and goes back through the chain of calculations ultimately to raw data: What surveys were handed out and what were the results in the survey? If you cannot audit the result, you do not know that it is correct."

In short, he said, it is only by sharing data—and "data" in the broad sense of everything that went into a conclusion—that the ideal of scientific knowledge being testable can be achieved. It goes back to the basic idea of science that people learn in high school or perhaps kindergarten, he said that its results flow from "publicly available, reproducible, everybody-can-stand-around-and-look-at-it data."

He said that in his field—risk analysis—there is a preference for seeing raw data. When members of professional societies who are engaged in risk analysis were asked, 69 percent said that it was very important to have access to the underlying raw data so that they can form their own conclusions. However, he added, only 36 percent reported that they typically had such access to raw data. Thus, "there is a perception of a big gap between what you would want to do a responsible job of coming to conclusions and what we currently have."

IMPROVING THE SCIENCE

Whatever the merits of reanalysis, there are other ways that sharing data does improve and advance the practice of science, Cox said. For example, he pointed out that if researchers know that other scientists may reanalyze their data, those original researchers are likely to do better science because of the possibility that someone will be looking over their shoulders. "Open access to key data and to models will encourage greater scrutiny," he said. "Other people will say, 'Let me see if I can come up with the same answer or ... suppose I make some other assumptions, would I get a very different answer? Let's go find out.'"

More careful scrutiny will encourage more careful research, Cox said, which in turn should result in a lower prevalence of false positives. The problem of false positives is a very serious one, he said, so more careful research would be a valuable benefit. Furthermore, more careful research will be accompanied by more careful interpretation of the data. If scientists refrain from overinterpreting the data—for example, by drawing causal conclusions from associations that are not necessarily causal—it will increase the trustworthiness of published results. "We

need greater trustworthiness in published results and in actions taken based on them. That is one class of benefits."

A related benefit, Cox said, is that the sharing of data lowers the "barriers to entry" to reanalysis or alternative analyses, which in turn encourages better and more frequent follow-ups. "We increase the value from the investments that have already been made in research," he said. "It is very costly to assemble a good database, so let's have a lot of people exploit that database to get as much juice out of it as possible."

Cox went on to suggest that researchers—particularly junior researchers—need to think of sharing data with others as part of science as usual. "Suppose a graduate student says to me, 'But, Tony, you don't understand. I plan to base my future career on mining [these] data that I have so painstakingly collected. I am certainly not going to share [them] with other people. There goes my career.' What I would say is, 'You have not paid the price of entry. The price of entry to being a real scientist is you do things in the open light. You do not monopolize them.'" He noted that researchers should expect that their results will be scrutinized and their data will be scrutinized. "That should be the price of entry to doing good science." He also suggested that concerns about advocates doing sloppy or skewed analysis to distort public policy are no reason to not share data. "I think that is part of the price of democracy. We should have people arguing about whether analysis is good. That is a perfect place to focus." He closed by suggesting, "Sharing data upon request should be the rule, not the exception. That should be the general expectation of good science."

THE BENEFITS OF MEGADATA

The benefit of having access to tens of thousands of data sets at one time was highlighted by George Daston, a Victor Mills Society Research Fellow at Procter & Gamble, in Session 4 of the workshop. Such "megadata" make it possible to do a variety of analyses that would not be feasible with the amounts of data available from one or a few data sets.

Daston introduced the topic by discussing the recent advances in information storage and analysis. "As a life scientist," he said, "I believe that we are entering the third great age of life science research—the first age being one of description and classification that started with Aristotle and went through Linnaeus and the second being reductionist, mechanistic biology [that led to] understanding at a deep molecular level of how things work and which started with Darwin and ended in about

2000." In this third age, biologists are beginning to put all of that reductionist information together in an effort to understand emergent properties. "This is the systems biology era," he said. "I think a lot of people are under the impression that the systems biology era has been driven by biotechnology—and it has been—but it has even more so been driven by the revolution in computational power."

In short, Daston said, the rapidly increasing power of computers has the potential to revolutionize environmental health and toxicology. In particular, he listed several old questions that increasing computational power may make it possible to answer:

- What is the relationship between chemical structure and toxicity?
- What does the universe of toxicity modes of action look like?
- How can we predict adverse outcome from initial molecular events?

But answering these questions will require huge amounts of data. "One of the things that we have done over the past 10 to 12 years," Daston said, "is to try to see what we can do on a large scale to make sense of all of the individual toxicology studies that have ever been conducted. What we have done is put together a relational database that can be searched by chemical structure or chemical substructure so that we can do things like propose an understanding of the toxicity of a new chemical based on analogy to chemicals that have already been tested and try to get a handle on the universe of modes of action or the universe of structural features of chemicals that convey some sort of hazard." Up to now, he said, these questions have not been answerable, and "because the boundaries of this are dark and fuzzy, we have always just thought there is no alternative other than to continue to test, and test chemical by chemical, and live with the uncertainties that they have."

Daston then displayed a slide that showed the numbers of traditional toxicology studies that he and his colleagues have assembled in their database, listed by the particular toxicity endpoint (see Box 3-1). For example, the database contains 69,000 studies of acute toxicity, 14,756 studies of carcinogenicity, and 11,923 studies of developmental and reproductive toxicity.

"You can see that there is actually a wealth of information out there," he said. "This is both from the published literature and what we can get out of the gray literature," with studies published in the gray literature (reports of research that has been performed by associations or companies but that

BOX 3-1
Taking Advantage of the Toxicology Literature

Size of toxicology database (number of traditional toxicology studies by endpoint):

- Acute toxicity (69,000)
- Genotoxicity (19,250)
- Subchronic toxicity (15,738)
- Carcinogenicity (14,756)
- Developmental/reproductive toxicity (11,923)
- Skin irritation (10,899)
- Skin sensitization (9,546)
- Eye irritation (6,811)

SOURCE: Daston, 2014.

has not gone through peer review) coming from various government agencies, such as studies that used data from a database of the U.S. Environmental Protection Agency (EPA) or the REACH database operated by the European Chemicals Agency. In total, the database contains nearly 160,000 studies.

Various things can be done with this type of information, Daston said. For instance, it is possible to put together a decision tree to help people decide whether a new chemical with a particular chemical structure is likely to have a certain type of toxicity. Daston showed an example of such a decision tree that he and his colleagues put together using data on about 800 developmental toxicants. The decision tree was largely based on chemical structural features, he said, but it could also be based on the putative modes of action for toxicity for reproductive or developmental toxicants.

"The simplicity of this is that for a toxicity as complicated as this one, we can break this down into essentially 25 groups of chemical features," he said. "What you can see is that with these computational approaches, we can start to get our arms around things that were not feasible before."

One thing that can be done is to determine whether, for a given chemical, there is any evidence in the literature that a chemical with a similar structure and chemistry is a developmental or reproductive toxicant. This is a valuable piece of information. "But more importantly," Daston said, "we can start to put together ontologies for modes of action of

toxicity, which organizes the field in a way that hasn't been done before." This in turn will allow regulatory agencies to start to understand which of the tens of thousands of chemicals about which there is relatively little information should be prioritized for additional study and which ones are probably not of great concern.

But this sort of approach demands a great deal of data to work, as do various other promising approaches, such as toxicogenomic analyses of modes of action. "These are really data-hungry exercises," Daston said. "For this kind of information, what we really need are data summaries," he said. "A lot of this is generated with a common protocol. The data are analyzed in essentially the same sorts of statistical ways. Summary statistics are probably the most important." It would be essentially impossible for him to do his work if he had to start with the raw data, he said. "Having gone through the process of compiling the database, [I can tell you] that this takes man-years worth of effort. Doing this ontology and doing this decision tree that you see here also take man-years worth of effort. That effort is only compounded if we would have had to start from raw data rather than some sort of summarized data."

Thus, he concluded, "For these purposes—and I think as a general principle—what you want to do is start with the most refined and processed data that you can use, that you can live with, because it saves you time. Think about what the question is."

Of course, he continued, there are indeed times when access to the raw data is important. "For other purposes, like genomic analyses, where we are still feeling our way through what the protocol should be for those things and how to analyze data, how to normalize data, in those sorts of situations, I think that raw data [are] probably also valuable to share, but probably not in exclusion to the refined data." And, indeed, he added, it has become the norm in his field to require the raw data to be submitted to a public repository before a paper is considered for publication. This is the case for most journals in his field, he said.

"I guess my point is that sharing of data—at least from where I sit—is more valuable than retaining the data," Daston concluded. "I realize that others have a different calculus from where they sit."

COMMENTS AND DISCUSSION

Several points related to the importance of data sharing were raised during the workshop. Workshop speakers and participants provided individual remarks that are summarized in this section.

Benefits of Data Sharing

During the discussion after Session 2, Dominici brought up some additional benefits of data sharing. "One of the main benefits [for] reproducible research is something that we saw happening by making the National Mortality, Morbidity, and Air Pollution Study fully reproducible," she said. Making the data behind the study accessible to other researchers elevated the level of discussion and the types of criticisms that were made about the study. Many of the smaller issues that other researchers had with the study were addressed just by making the data available on a website for those researchers to download and study, and once they had those data, those other researchers could interact with the scientists behind the study on a much more substantive basis. "That immediately elevated the scientific rigor of the discussions that we were having with other investigators," Dominici said. "I think that that is one of the main benefits because then we are talking about important issues about whether or not these results are reproducible and valid. Clearly, we are not analyzing a two-by-two table. We are analyzing a tremendous amount of data."

Another advantage of sharing data, Dominici said, is that it keeps researchers from having to "reinvent the wheel"—that is, from having to repeat the work that previous investigators have already done. "It seems silly to me that another investigator is going to start from what I started at 6 years ago or 10 years ago when he or she can start from where I left off," she said. With all of the technology in place to share data among researchers, she said, such sharing should really increase the speed of scientific discovery.

In her opening and stage-setting remarks, Lynn Goldman, dean of the Milken Institute School of Public Health at George Washington University, offered another scientific benefit of sharing of data: it can be crucial in carrying out systematic reviews in a particular field. Often, she said, in a systematic review it is necessary to reanalyze at least some of the data in the studies being reviewed, and that is possible only if the original researchers make their data available to other scientists.

Gwen Collman, director of the Division of Extramural Research and Training at the National Institute of Environmental Health Sciences, described another class of benefits. "One of the benefits of sharing data broadly is that you can bring more intellectual power and people from different disciplines and different perspectives into analyzing data," she said. "We are not only looking at data to redo or rethink or refute or

confirm somebody's analysis, but with so [many] data being generated so quickly and in the hands of so few, there really are a lot of people who could be brought into the whole scientific enterprise. New discoveries and new directions can come from looking at data in fresh and new ways. I do not want us to lose the thought of that as we continue our discussions for the rest of the day because I think it is a very important part of data sharing."

When Is It Valuable to Have Access to Raw Data?

During the discussion after Session 3 of the workshop, Goldman described a situation in which it is valuable to have access to the raw data. At one point in her career, she said, she was an assistant administrator at EPA in charge of the office that regulates chemicals and pesticides. The agency receives the raw data underlying studies of toxicity and reviews those data, and she found that there were always interesting things to be learned from the raw data. "What it has to do with sometimes are observations that are recorded in the margins about things that are going on with the animals that then have led to new discoveries about some toxicities," she said.

In particular, she spoke of an experience with the fungicide vinclozolin. The Organisation for Economic Co-operation and Development (OECD) testing protocol had not been looking for antiandrogen effects from the fungicide, but the observations that were recorded indicated that it had the potential for such effects, which led to laboratory investigations that were quite different from the original OECD studies. "And none of that would have been detected if EPA scientists had not been looking at the raw data," she said, "because, of course, the contract lab for the company was not looking for things like that. And if EPA had only had the data summaries, none of that would have been discovered." In short, it can sometimes be quite valuable for the scientific community to have access to the raw data.

Documenting Data Sharing

Dominici noted during the discussion after Session 2 of the workshop that asking or requiring investigators to make their data—and often their software—available to other researchers puts a tremendous burden on the investigators. Therefore, she said, some sort of incentive should be provided to those investigators, particularly junior faculty members.

For example, she said, "if you publish a paper that is fully reproducible, that could be an additional comment [in the journal of publication], and it could be something that really improves and plays an important part in your career. Deans, chairs of the department, journal editors, government agencies—if they are all united, they could definitely make this possible with the right incentives, which right now are not in place."

Bernard Lo, president and chief executive officer of The Greenwall Foundation, had some similar thoughts. He suggested that a researcher's standing as a scientist and, specifically, as a faculty member should depend not just on gathering data in innovative ways and developing new ways of analyzing them but also on sharing these data in a way that generates even more knowledge. It would be necessary to figure out a way to track that, he acknowledged. "It is not just how many people use your data, but how many really interesting additional studies came out of them." There should be metrics not just for sharing data but for sharing data in really positive, creative ways that lead to new science, he said. He explained that perhaps this kind of productive data sharing should be an expectation in faculty promotion reviews.

REFERENCES

Daston, G. 2014. *Data uses and data requirements*. Presentation at the workshop Principles and Obstacles for Sharing Data from Environmental Health Research, Washington, DC.

Ioannidis, J. P. A. 2005. Why most published research findings are false. *PLoS Medicine* 2(8):e124.

Lehrer, J. 2011. Trials and errors: Why science is failing us. Wired. Available at http://www.wired.com/2011/12/ff_causation/all/1 (accessed October 27, 2015).

Sarewitz, D. 2012. Beware the creeping cracks of bias. *Nature* 485(7397):149.

4

Issues and Challenges Associated with Data Sharing

This chapter summarizes presentations on a number of challenges associated with the sharing of data, including obstacles to releasing data, privacy and confidentiality problems, and informed-consent issues. The discussions concerning these issues can be found in the last section of the chapter.

OBSTACLES TO THE RELEASE OF DATA

There are a number of different obstacles associated with the release of data. Some are related to concerns of the scientists who have generated the data, some are related to the concerns of businesses or other organizations that have paid for the collection of the data, and some are practical issues having to do with the administration of the data.

Concerns About Adversarial Science

It is, unfortunately, the case that the science surrounding environmental health issues can sometimes become very contentious and adversarial. As was the case with tobacco companies and the research showing a link between smoking and lung cancer, companies that profit from the production or use of certain chemicals or products can resist research that indicates that those chemicals have health risks. And this can put researchers in a bind, said Daniel Greenbaum, president of the Health Effects Institute (HEI), during his presentation in Session 4 of the workshop. “There is an inherent tension between collaborative scientific data sharing and what often can be adversarial science,” he said. “The challenge here is that investigators invested time. They probably should

be able to make their data available, but they are being engaged by advocates in controversial cases who are not there really to advance knowledge. Their primary purpose is in undermining the study." If the investigators believed that these advocates were interested in advancing scientific knowledge, the investigators would be willing to share, but believing that the advocates are only interested in looking for weaknesses in the data that can be used to discredit the researchers makes them hesitant to make the data public.

In a worst-case scenario, the original researchers might even find themselves accused of scientific fraud by such advocates. This happened in California not too long ago, Greenbaum said. The case was meritless and "got thrown out," he said, but it was still quite unpleasant for the investigators who were accused. "You can imagine—that is a pretty chilling activity."

As a result of such adversarial tenseness, Greenbaum said, "the new investigators, many of whom are talented biostatisticians and epidemiologists, get frustrated because they cannot even call up the original investigators once they have the data" and ask questions about the data, such as details about how they were generated. "The original investigators do not want to talk to them."

With the sharing of data encumbered in this way and the usual open scientific dialogue closed down somewhat, Greenbaum explained that it becomes much harder to advance the science in these areas—not just replicating the initial work but also extending it and carrying out new analyses on the data sets. All of these valuable outcomes of data sharing become the victims of distrust and suspicions.

Tensions Between Researchers and Opponents

To reduce some of the problems surrounding the sharing of data, Greenbaum suggested a different approach: finding ways to ease tensions between the researchers who collect the data and their opponents who are interested in discrediting the data or the conclusions that have been drawn from them. "Is there a way to facilitate a true dialogue," he asked, "between those who fund the adversarial reanalysis—and I am not limiting that only to industry, ... although it tends to happen more with industry—and the scientific community that can set a foundation for more thoughtful and independent testing of key studies? Is there a way to do that in a way that would make sense and that would start to lower the

decibel volume around this incredible distrust ... between the two parties here?"

There may well be no bigger issue in the broad discussion about data sharing than how to ease tensions between these two opposing camps, Greenbaum said. "There are lots of issues around the technical and the other stuff, but it is this issue of how to lower the temperature of these discussions that allows science to shine through on the other end."

If researchers are to be comfortable sharing sensitive data with others, they must have ways to make sure that the confidentiality of individuals in a data set can be preserved when the data set is shared. This can be particularly challenging for environmental health research, Greenbaum noted. "To do a good air quality study and estimate exposure, you do need to know where somebody lives," he said. "And once you know where somebody lives and that they have died on a certain date, it is pretty easy to figure out who they are."

EPA has asked the HEI for advice on how to share confidential data without compromising the confidentiality, Greenbaum said, and, according to the institute, there are basically three options for doing that.

The first option, which is very common, is to have full collaborative data sharing with original investigators. "If you go to the American Cancer Society website, you can apply to them and ask to collaborate with them," Greenbaum said. "There are many other mechanisms for doing that."

The second option is the sharing of data through data use agreements. The National Institutes of Health (NIH) has a number of policies for doing this, Greenbaum said. "This is essentially what we did with our own HEI reanalysis. This provides the new investigators much greater access to the underlying data, but with a very significant requirement that they will protect any personal information from disclosure. I do not think I knew it was a $250,000 fine [for failing to protect any personal information] when I signed the data use agreement for the Harvard Six Cities Study and the American Cancer Study, but I knew it was significant penalties, and we pay attention to it. That is a mechanism that is out there, and it is used on a regular basis."

The third option, which has become increasingly popular in recent years, is to share deidentified data files. "It is possible to do that, for example, for the air quality data sets for the cohort populations," Greenbaum said, "but the reality is you couldn't even properly replicate those studies today ... because a deidentified one would not allow you to have location and a variety of other things that are absolutely essential."

Similarly, it would be impossible to do any sort of reanalysis or additional analysis of the data without having the location information and other types of information that could allow reidentification of the subjects. "You really have to have that kind of information," he said.

A variation on the second option that Greenbaum mentioned—the sharing of data through data use agreements—is the use of secure data enclaves for the sharing of data. In this setup, sensitive data are kept in a particular physical location, and researchers use the data there so that the data themselves never leave the secure enclave.

Business Considerations Related to Data Sharing

Businesses must take into account various issues when deciding whether to share their data. Two of the main ones are concerns about opening themselves up to liability and other costs and worries about losing the value of confidential business information.

During Session 5 of the workshop, Joseph Rodricks, principal of ENVIRON, discussed his experience with businesses whose products might have health effects. "The concern in this country about product liability and toxic tort litigation is very high," he said. "When something gets elevated to a known human 'something,' litigation breaks out all over. That is the experience of most companies. You might have at least some understanding why they might be concerned about such a finding."

In particular, Rodricks spoke about how businesses react when studies appear that seem to have a strong potential of adversely affecting their business interests. "The ones of most concern ... are those emerging from human studies where authors are looking at causal relationships between some exposure and a disease," he said. "Industry gets [really] nervous when studies like that appear if it is their product." They are concerned about increased regulation as well as about liability.

"I find companies have different attitudes about what they want to do in those circumstances," Rodricks said. "I discuss a lot of possibilities with them, including doing additional studies. I have heard from some companies that they want to find a way to undermine a study. I just say, 'Go away. I will not engage in that kind of activity.'"

He agreed with Greenbaum's concerns about adversarial science. "I have witnessed these very adversarial confrontations, and they are very difficult to even watch," he said. "I certainly do not want to get involved in them."

Rodricks noted that the process could be improved by beginning to assemble some guidelines on best practices for sharing data from environmental health research. These guidelines could help people organize their data management process so that some data are prepared for sharing when needed. He pointed out that a document from the Oak Ridge National Laboratory on best practices for preparing environmental data sets to be shared and archived has excellent detail on what investigators can be doing to prepare to publicly release the data as they develop them.[1]

The Business Value of Data

During Session 3 of the workshop, Greg Bond from the Dow Masters Fellowship Program at the University of Michigan spoke about some of the other business considerations related to data sharing. He noted that any raw data that industry generates to support a product registration are accessible to the relevant government authorities under a variety of statutes in the United States and also abroad. "The vast majority of these data will have been from animal toxicology studies that were conducted under good laboratory practices," he said. "I can personally attest, as someone who ran a large, in-house toxicology laboratory, that the U.S. EPA comes in unannounced and will audit the data, all the raw data. And as a consequence, we have to keep the tissue blocks and the slides for 35-plus years so that they can audit again in the future, if necessary."

Regulatory agencies are generally obligated to protect the commercial value of the data collected by companies. The key fact, Bond said, is that the data generated by industry researchers have commercial value. "This is recognized under FIFRA [the Federal Insecticide, Fungicide, and Rodenticide Act] and similar pesticide regulations globally," he said. "Each of those regulations provides for exclusive use of the data by those who bore the cost of generating the data and for fair and just compensation [for] them by others who want to use the same data to achieve their own registrations and access to the market. The goal here is to prevent free riders and preserve the incentive for companies to invest in innovation and research."

[1] "Best Practices for Preparing Environmental Data Sets to Share and Archive" is available at http://daac.ornl.gov/PI/BestPractices-2010.pdf (accessed February 23, 2016).

However, under pesticide regulations there are always time limits on how long a company can be compensated for those data. They are usually in the range of 5 to 15 years, Bond said. "That data compensation is becoming even more important as the rise of low-cost competitors, particularly from Asia and Eastern Europe, has increased," he said. "These competitors already have some built-in advantages, and, frankly, we do not need to subsidize them by giving our data away to them."

In particular, he said, he saw firsthand the sorts of advantages that these competitors have over U.S. companies when he spent 3 years working in China and observed the difference between how American companies and Chinese companies dealt with environmental health and safety (EH&S) concerns. "Our folks would wonder why it cost us $30 million to build a plant in China when the local producers could build it for $10 million," he said. "A lot of that [was] EH&S protections that were built in."

Thus, it makes sense, Bond said, that the companies that pay to collect environmental data consider those data to be business assets and are generally not eager to share them without compensation. In particular, data may offer the company that collected them some competitive advantage over their competitors—some particular knowledge that their competitors do not have or some insights that may lead to solutions that the competitors might not come up with. As a result, companies often treat environmental data as confidential business information and try to protect them from their competitors. The preservation of such confidential business information can come into conflict with the imperative to share research data, and finding the proper balance between the need to keep business information confidential and the desire to share scientific data can be difficult.

One way that businesses deal with the conflict, Bond said, is to provide higher-level information that is based upon the data. "Certainly, we believe that information should be publicly accessible that is sufficient for people to take action," he said. For example, companies provide product labels and material safety data sheets with summary information that is sufficient to allow people to determine the steps that they should take, if necessary, to protect themselves. It may be necessary in some cases to improve the quality and quantity of that information, he said, but still, the company would be supplying information as opposed to the underlying data that the information was based on.

Bond explained what happens when someone wishes to question the conclusions of a study that, for instance, was submitted in support of a

position or a rule that EPA is considering. In that case, it would not be enough to look at higher-level information based on the data—the original data themselves would be required. First, Bond said, EPA itself always has access to the data, so it can check whether the conclusions of the study did indeed follow from the data. And, indeed, EPA generally does audit the data. Beyond that, there may be a time limit beyond which the company can keep the data, particularly toxicology data, confidential, and beyond that time limit the data are available. However, he added, he could not think of a single instance when someone asked for access to the data underlying a particular toxicology study. "It has not been an issue that we have had to deal with very often."

Finally, Bond said, there is the issue of who pays for access to data that a company has paid to collect. The companies that collected data deserve and expect compensation for supplying those data to others, but where should that compensation come from? If another company wishes to use the data, the answer is clear, but if, say, academic researchers are looking to use the data, the answer is not so clear.

Administrative Issues

In addition to scientists' concerns and companies' concerns, data sharing is also complicated by various administrative issues, Bond said. For example, one issue that arises on occasion is the question of who owns the data. This is particularly an issue when data have been collected by groups of companies or institutions. A reasonable number of toxicology studies have been done by industry consortia, for instance. These were groups of competitors who came together years ago to share the cost of conducting the fundamental toxicology research, and over the years some of those companies have been absorbed into other companies via mergers and acquisitions. In a few cases, such absorptions have happened multiple times. And in some cases the companies involved in one of the consortia have gone out of business. The result, Bond said, is that it can be very difficult to sort all of this out to find someone who can make a decision about data sharing.

"As you can imagine," Bond said, "not all competitors play nicely together. Some even resort to gamesmanship to try to exclude competitors from the market. Things can get nasty and messy in a hurry in these discussions."

According to Bond, among the other administrative challenges is the fact that organizational policies and procedures and logistical issues may

differ from organization to organization. For example, institutional review boards at different universities may have different rules governing the procedures for sharing data.

How Will Data Be Made Available and Who Will Pay for Data Sharing?

Jerry Blancato, director of the Office of Science and Information Management at EPA's Office of Research and Development, discussed governance and budget issues during his remarks in Session 5 of the workshop. He noted that multiple people talked about the need for definitions during the workshop, but from the point of view of someone who is in charge of the government asset of data, the question really is "where are the data?" Then decision makers may ask, how does one make the data publicly available, are the data readable by today's machines, how should archived data be handled, and how should one plan for future data requests 20 or 30 years from now.

One major obstacle to extensive data sharing is its cost. "This is not a zero sum game," said Blancato during his presentation. "This is going to cost us. It is going to cost us in dollars. There is no question about it."

"The problem is that the budgets are not expanding to cover this," he explained. "A balance has to be made." If data are going to be made publicly available, then the money to do that is going to have to come from somewhere, perhaps from direct research support.

"One could make the argument ... that from a societal point of view, sharing the data will have far greater benefits," he said. "But we have to measure that. And, as someone else said, we have to do some work to communicate that and convince the community—both the research community and the public community—that that in fact is true.... We are being screamed at from Congress, from the White House, from internal forces, from external forces. 'Make the data available. Why can't my data get out there so people can see it?'" But the money has to come from somewhere, and right now there is no agreement on where the money will come from.

Blancato noted that governance is key to figuring out how the data will be made available. Technology may be able to help formulate a solution to the sharing of data publicly, but pieces will still need to be protected and not shared. It is important to take advantage of partnerships with industry experts in Silicon Valley and elsewhere who know how to reasonably store tremendous amounts of data relatively cheaply. He

reinforced the importance of predicting and utilizing upcoming technology in planning for the future so that the data can be used and read 5, 10, or 30 years from now.

PRIVACY AND CONFIDENTIALITY ISSUES

One of the obstacles to the release of environmental data is how to ensure, as much as possible, the privacy of the individuals whose data have been collected, as some of these data, such as medical history data or employment data, can be quite sensitive. These individuals were often promised confidentiality before they agreed to provide those data. This section summarizes presentations that focused on the risk that participant confidentiality could be breached in some way or another.

What Is the Empirical Knowledge About Reidentification in Data Sets?

Concerns about reidentification are causing some organizations to pause in their plans to share data, said Julia Brody, executive director of the Silent Spring Institute, during her presentation in Session 2 of the workshop. For example, she indicated, the institute would like to share its data set through EPA's ExpoCast, an online data resource, but there are questions around whether the data may be reidentifiable, especially when linking air pollution data and personal information, such as household characteristics and consumer product purchases. These questions led Brody and her colleagues to establish a partnership, with funding from the National Institute of Environmental Health Sciences (NIEHS), to empirically investigate the sources of privacy risk in environmental health studies.

"We have looked at 11 major environmental health studies that have collected personal-level exposure measurements to see what kind of data they include that might be vulnerable," Brody said. "We are focusing particularly on data that might be linked to real estate databases, property transfers, tax records, [and] zoning records and data that might linkable to professional licensing registries, for example, pesticide applicators in the Agricultural Health Study or teachers in the California Teachers Study." The project is also looking at a category of data that includes such things as what a neighbor might know about a person: what kind of fuel is used in a person's house, how many pets a person has, when a house was remodeled, and so on.

One goal of the study is to empirically evaluate the reidentifiability of data that the Silent Spring Institute had collected by doing a reidentification experiment to see how many participants could be correctly reidentified by linking study data and public data sets. The institutional review boards that reviewed the original studies have approved this, but the Massachusetts Cancer Registry has not. "We are still talking to them about it," she said. "In order to generate empirical data about this question, we will need to continue that discussion with them and hope we reach a resolution."

Another goal of the study is to find out how study participants think about privacy issues. "We interviewed participants in the Personal Genome Project, which is an open-consent study in which participants go through a training process and post their data on the Web," Brody said. "These participants are not representative of people in studies, but they are really on the frontier of giving up their privacy for science and for public health benefit."

The study hopes to answer three questions, Brody said. "One, what is our empirical knowledge about what could lead to reidentification in our data sets? Two, what are the potential technical solutions to that in terms of masking or synthesizing data or creating server systems that would respond to queries? And three, what can and should we promise to our study participants in the informed consent?"

A key question is how likely is reidentification of subjects in existing data sets that have had the obvious personal identifying information removed. Two presenters offered very different answers to that question that are summarized in the next section.

Risks of Participants Being Identified

One of the standard ways to share sensitive data in which the privacy of the subjects must be respected is to "deidentify" the data before they are released.[2] However, Daniel Barth-Jones, assistant professor of clinical epidemiology at Columbia University's Mailman School of Public Health, pointed out during his presentation during Session 3 of the workshop that removing identifying data from a data set decreases their usefulness.

He focused specifically on the value of sharing deidentified data, or data stripped of names and other information, such as quasi-identifiers

[2] There is not a precise definition of deidentified data, and much debate and uncertainty about what constitutes deidentified data remain.

which in combination could identify the people who supplied the information in a data set. "I believe that there is a great societal value to deidentified data," he said, "that it provides invaluable public good, and that it is an essential tool for our society in supporting scientific innovation and research. It helps drive forward innumerable scientific and health research advances. It greatly benefits our society as a whole and yet still provides strong privacy protections for individuals."

"As we move towards expanded health information technology and electronic medical records," he concluded, "it will yield even more deidentified clinical data, which I believe will support important advances in health science."

"The inconvenient truth is that we are stuck with a trade-off," he said. "When we deidentify data, we necessarily degrade that data in different ways, either through overt information loss, like restricting dates that are more specific than a year in HIPAA [the Health Insurance Portability and Accountability Act], or through the trade-off of grouping people together, for example, so that they are not observable as specific individuals" (see Figure 4-1). The less information that is available about the individuals in a data set, the less that can be determined about how different variables are related to environmental exposures. For instance, he said, if information about race is removed from a data set to make it more difficult to identify the subjects, then it becomes impossible to determine if a particular environmental hazard affects members of one race more than another. Furthermore, removing many of the identifying data does not completely guarantee that the subjects cannot be "reidentified"—that is, have their identities deduced from the remaining information, such as sex, age, race, and general location (e.g., zip code or Census tract).

Point: The Risks of Reidentification Are Relatively Low

Barth-Jones stressed in his presentation that much of the current discussion about privacy policies has been driven mainly by anecdotes and, in particular, by a few dramatic accounts of reidentification. However, Barth-Jones said, instead of basing policy on a few dramatic examples like this, it is important to look at the entire body of evidence for a group of people and determine what the average risks are. "It is important to know what the vulnerabilities are," he said, "but if we do not

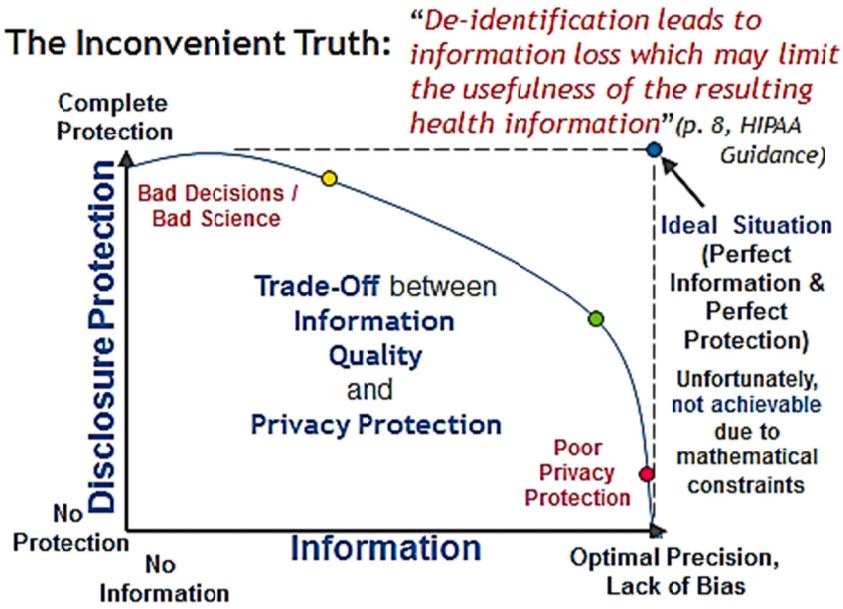

FIGURE 4-1 The trade-off between privacy and scientific usefulness of data.
NOTES: HIPAA is the federal Health Insurance Portability and Accountability Act of 1996. The primary goal of the law is to make it easier for people to keep health insurance, protect the confidentiality and security of health care information and help the health care industry control administrative costs. The HIPAA Privacy Rule provides federal protections for individually identifiable health information held by covered entities and their business associates and gives patients an array of rights with respect to that information. At the same time, the Privacy Rule is balanced so that it permits the disclosure of health information needed for patient care and other important purposes.
SOURCE: Barth-Jones, 2014.

get down to the specifics of what the overall risks are, we will not be able to make good policy decisions."

In particular, Barth-Jones said, he was focusing on the issue of reidentifying patients from data sets that have been deidentified according to HIPAA protocols. These data are collected by doctors and other health care professionals during the provision of health care, and they can be made available for scientific research if they are deidentified as specified by HIPAA.

"It has been claimed that deidentification does not work," Barth-Jones said, "and I think in the pre-HIPAA context, that is probably the

case." In particular, he pointed to recent work by a fellow workshop participant, Latanya Sweeney, professor of government and technology in residence at Harvard University, who examined deidentified data and various data sets and was able to use such information as birth date, gender, and zip code to reidentify a significant percentage of the individuals in the data sets—from 27 percent of individuals in the Personal Genome Project to as many as 87 percent in an earlier study. However, he said, things are different now under HIPAA.

"The reality is that HIPAA-compliant deidentification provides important privacy protections," he said. "The estimate for the 'safe harbor' deidentification[3] was that about 4 in 10,000 individuals might be reidentifiable." Barth-Jones mentioned that in a systematic review by El Emam and colleagues (2011), of the 14 reidentification attacks that they surveyed, only 2 were done on data sets that had been deidentified according to current standards, and only one of those data sets contained health data. In that case, the reidentification attempt managed to identify only 2 out of the 15,000 individuals whose data were in the data set. "I think at that point we have to question why someone would go through the effort to do that," he said.

On the other hand, though, there is no such thing as perfect and permanent deidentification. And because the future will always bring new ways to reidentify data, it is impossible to guarantee individuals that their data will never be identified. It is possible to make reidentification difficult, but not impossible, he said.

There are various approaches to deidentifying data, each with its own advantages and disadvantages, Barth-Jones noted. For example, the "expert determination" method is a little bit more flexible than the "safe harbor" approach.[4] "It helps us balance the competing goals of privacy protection and preserving the utility and statistical accuracy of deidentified data," he said.

In thinking about the deidentification and possible reidentification of data, there are two things that are important to understand, Barth-Jones said. "One is this myth that we can build a perfect population register." It is difficult to develop a population register that is comprehensive, that is,

[3] The safe harbor approach requires removal of all 18 listed identifiers under the HIPAA Privacy Rule.

[4] The expert determination method can be used when it is preferable to not remove all 18 listed identifiers. The method requires confirmation from a qualified statistician that the risk of identification is very small on the basis of the retained identifiers.

that includes everyone. For instance, many population registers depend on voter registration lists, but only about 70 percent of Americans are registered to vote.

The other thing to keep in mind is that various errors and inconsistencies will inevitably arise in linking data between the sample and the population, creating what Barth-Jones referred to as "data divergence." The basic approach to reidentification is to match variables from a population data set—a list of registered voters, for example—with variables from the sample data set under investigation. But there are various ways that the data in the population data set can diverge from the data in the sample data set. First, the values of the data variables can change over time. People move from zip code to zip code, their income bracket changes, or their marital status changes. All of these things make it more difficult to draw a link between the sample in question and the overall population. There are also missing and incomplete data as well as keystroke and other coding errors in any data set, which, again, will make it harder to link people in the sample data set with individuals in the population data set.

A fundamental problem with efforts to protect individuals from identification, Barth-Jones said, is that as more is done to deidentify or anonymize the data, the less useful the information is for statistical analyses. To illustrate, Barth-Jones showed an original analysis that clearly showed a race-based effect: as depicted in Figure 4-2, African American, Hispanic, and Asian groups all scored differently from white Americans (the leftmost group). But when the various racial groupings other than white were combined to anonymize the data and make it harder to identify individuals, the effect disappeared, and all that remained was a white group and an "other" group (the rightmost group) between which there were no statistically significant differences.

One important thing to keep in mind when trying to deidentify data, Barth-Jones said, is the difference between whether someone is unique in the sample under consideration or unique in the larger population. "If they are unique in the sample, we call them 'sample unique.' If they are unique in the larger population, we call them 'population unique.'" In general, in any given sample there may be individuals with unique identifying information; for example, there may be only one person in the sample with a particular age, sex, location, and degree of education, but in the larger population there may be several people with that particular set of identifiers. "It is really only those records that are unique

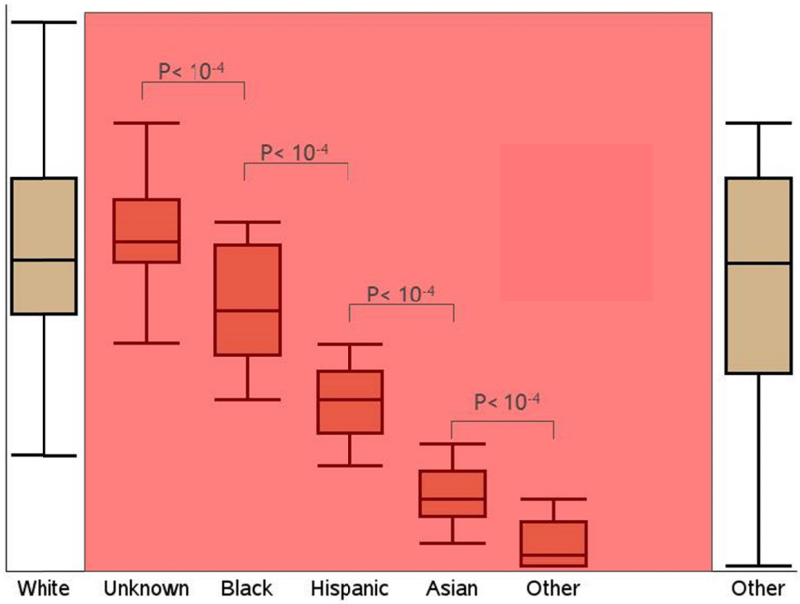

FIGURE 4-2 K-anonymity can hide heterogeneities.
NOTE: Released data provide k-anonymity if the data for each person cannot be distinguished from at least k-1 individuals whose data also appears in the release.
SOURCE: Barth-Jones, 2014.

in the sample and in the population that are at a clear risk of being reidentified using exact record linkage," he explained.

Unfortunately, he said, efforts to deidentify samples sometimes do not take into account the differences between sample uniqueness and population uniqueness. Some have suggested that deidentification methods collapse variables to the point that only a tiny percentage of the sample is unique. This is generally not necessary because the real question is often not how many records are sample unique but rather how many of those sample-unique records are also population unique. When people ignore this distinction and try to minimize the sample-unique records, it can cause the data set to lose most of its value for understanding environmental effects on health. "If we are going to do this to our statistical analyses," he said, "we might as well all give up and go home."

Barth-Jones noted that it is important to obtain a better understanding of the risks that deidentified patient data can be reidentified to improve patient confidentiality. "How do we move beyond anecdotes to a

rigorous scientific evidence-based risk management approach for dealing with reidentification risks?" he asked. "I would suggest we want to do that using quantitative policy analyses. These have been used for decades by agencies like the EPA and the Energy Department to address difficult risk management questions. There is a lot of uncertainty in reidentification risk assessment, and I think we can use these methods here."

Counterpoint: We Need a New Way of Thinking About the Risks of Reidentification

At various points in the workshop, both Barth-Jones and Sweeney referred to perhaps the best known case of reidentification, which occurred in 1997 when Sweeney, then a graduate student, was able to pick out from a state insurance data set Massachusetts Governor William Weld, whose data had been deidentified. The fact that she was able to identify Weld, even though the data set had been stripped of anything that directly identified the people in it, seemed to be a striking example of how easy it is to circumvent deidentification. Sweeney offered her own thoughts on reidentification during her presentation in Session 5 of the workshop. Work on reidentification over the past 15 years has resulted in a new way of thinking about the risks of reidentification. She discussed how reidentification experiments are conducted—for example, by computer science scholars—to evaluate the privacy risks of peer-reviewed and published data sets so the reidentification methods and results can inform privacy protection policies and strategies.

The possibility that data will be reidentified scares many people, Sweeney said. Scientists worry that if the subjects whose data are in their data sets are reidentified, it will discredit them and perhaps make it more difficult to attract subjects in the future. But as a computer scientist, she said, she thinks of reidentification in a very different way.

"[R]eidentification is a sufficient and necessary condition for improving privacy protection," she said. Without real-world demonstrations of data reidentification that indicate actual vulnerabilities and risks, there will be bad policy and little scientific progress in privacy-enhancing technologies. To zero in on the appropriate types of privacy protections—whether they are stronger than today's protections, weaker, or, more likely, totally different from what is now known—it is necessary to go through cycles of data being reidentified, protections being changed in response, and so on.

"Computer science had this exact same problem many years ago in encryption," she said. "It used to be the case that some people would rely

on encryption for national security. Somebody would break it. They would publish it. The national security people would freak out. Businesses would freak out. Somebody else would come up with a better system, and somebody else would break that. They created this cycle until eventually we got the strong encryption that we use today. Today, I can tell you exactly how your credit card is encrypted over the Internet, and even knowing how it is encrypted does not make it possible for you to break it. We could have never gotten there if we did not have those cycles."

Sweeney described current thinking on reidentification as being like an argument between two lawyers who have different political positions. "One camp says everything can be reidentified, and the other camp says nothing can be reidentified. And neither one of them is right. But now how do we clarify it?"

The current situation, Sweeney argued, is the result of 15 years of talking about the risks of reidentification with very few people actually looking at the details of how reidentifications occur and how they can be prevented. For example, her work in reidentifying William Weld from the Group Insurance Commission in Massachusetts is well-known. "It is cited in the HIPAA privacy regulation preamble," she said. "It is cited in preambles in other countries' privacy regulations." But although she has made more than 20 attempts to publish a paper describing the details, it has never been published. "Not because I did not try," she said. "Not because I did not write the papers, but because, in general, reidentifications are disruptive. Despite a significant history after all those attempts, we do not have a history to turn to."

The next large-scale reidentification she did was a case in southern Illinois for the Department of Public Health. She reidentified children in a cancer registry. Her results were scored, and she was told that she was absolutely accurate in 20 out of 22 of the cases, she said. "The judge saw the results and sealed the case and said you may never tell anyone how you did that."

She did reidentifications in New York City and for the U.S. Census Bureau. In both of those cases, she found that the data were vulnerable to reidentifications even as the data "were already going out the door." In each case, she said, she was asked to give the organizations time to figure out how to deal with the vulnerabilities. "I did," she said, "and for years there was nothing done. They just continued to ship the data in the same way as if the vulnerability never existed, and the papers did not either."

The bottom line, she said, is that it is necessary to study situations in careful detail—including what information is available and who has access to it—if one is to understand the true risks of reidentification.

She is now working at assembling that evidence, she said. As part of that effort she has put together a map of how health information travels between organizations and individuals (see Figure 4-3). "This is what the data map looks like today," she said. "You can visit it on thedatamap.org. Every node, if you click on it, will list names of entities and document how we know they engage in that sharing practice." The entities in bold font are those that are not covered by the privacy rules in HIPAA—and they are surprisingly numerous.

One of the major nodes in the map is hospital discharge data, she said. "There are about 33 states that share the [deidentified] data. Only three states use HIPAA standards. The other 30 states use standards that are weaker than HIPAA." When she used Freedom of Information Act (FOIA) requests to find out who is getting these data, she found out that researchers are nowhere near the top 50 recipients on the list.

Sweeney showed a sample of deidentified hospital discharge data that she bought for $50 from Washington State, which she then linked to newspaper articles that appeared during the same time period. "We stick with news stories," she explained, "because news stories have the same kind of information that an employer knows about you, a creditor knows about you, your family and friends know about you, and you could link the data as well."

She linked those data to publicly available data from various lists. "We do not use voter lists anymore," she said. "We use public records because they are readily available. They are far more comprehensive. Everybody is in it. Your nicknames are in it. Your cell phones are in it." From that publicly available information she was able to get zip codes for individuals, which she linked with the hospital data to identify people who had been discharged from the hospital.

"We found that we had accurately reidentified 45 percent," Sweeney said. They determined that accuracy rate by giving all of the names to Bloomberg News, and the publishers of Bloomberg News and Jordan Robertson, the reporter, contacted each person on the list under an agreement that only if the person agreed would they actually release the name.

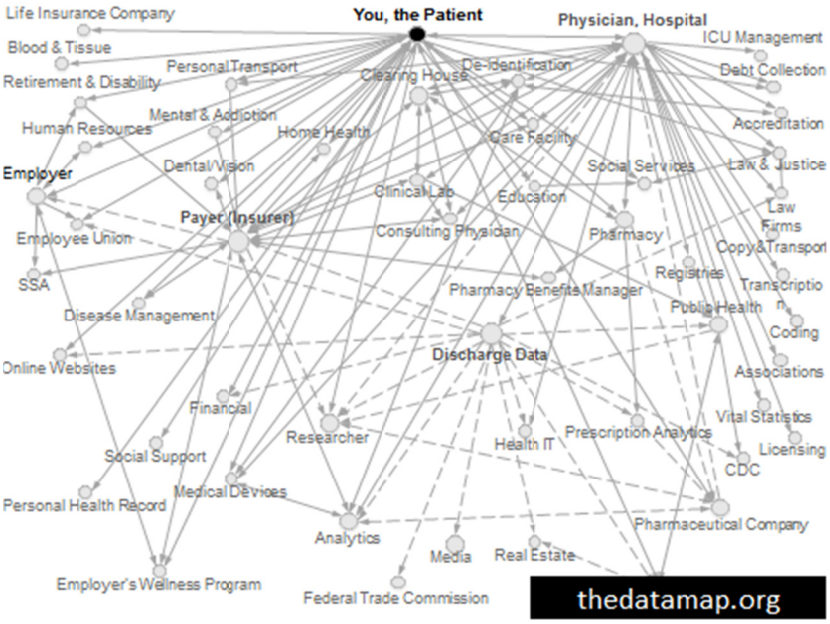

FIGURE 4-3 Where does health information travel?
NOTE: CDC = Centers for Disease Control and Prevention; ICU = intensive care unit; IT = information technology; SSA = Social Security Administration.
SOURCE: Sweeney, 2014.

This is the reidentification problem in 2014, she said. There is so much information floating around that is readily available to anyone who knows where to look that it is very difficult to get a handle on the exact risks of reidentification. But the first step in understanding those risks is to get a clear picture of exactly what information is available and how the different sources of information can be combined to identify people who appear in data sets that are supposed to be deidentified.

INFORMED-CONSENT ISSUES

Before an individual can take part in a scientific or medical research study, it is generally necessary for that person to provide "informed consent." That is, the person must be told enough about the study that he or she develops a reasonable understanding of the risks of participating and then still consents to take part. In universities, governmental agencies, and other organizations, the process of getting informed

consent—including details on how the specifics of the study are presented—is overseen by an institutional review board, whose purpose is to make sure that subjects in studies are not put at unnecessary risk and that they understand whatever risks might be involved.

Kevin Casey, the associate vice president for public affairs and communications at Harvard University, offered a detailed example of how informed-consent issues arise and affect scientific research during his presentation during Session 3 of the workshop. Researchers at Harvard followed a large number of subjects in six cities for nearly two decades as a way of studying the relationship between exposure to air pollution and mortality risk. The so-called Six Cities Study,[5] data from which were first published in 1993 (Dockery et al., 1993), played a major role in the establishment of air quality standards.

According to Casey, the informed-consent form signed by participants in the study said, "Harvard University School of Public Health hereby gives the assurance that your identity and your relationship to any information obtained by reason of your participation in this study of respiratory symptoms will be kept confidential and will not otherwise be disclosed except as specifically authorized by you." In short, Casey said, the subjects were assured that their data would remain private and that the data would remain in the hands of Harvard University.

The data for the Six Cities Study were of three types: some data consisted of basic information concerning the individuals and their health histories, some data described the subjects' respiratory systems when they were admitted to hospitals and where they were living at the time, and the third type of data consisted of death records. The location data were important, Casey noted. "That is different than other kinds of biological issues, where, if someone is eating broccoli, you do not care where they are located... . In this instance, it is important to know where

[5] The Harvard Six Cities Study evaluated a well-characterized cohort of adults from six communities (Harriman, Tennessee; Portage, Wisconsin; St. Louis, Missouri; Steubenville, Ohio; Topeka, Kansas; and Watertown, Massachusetts) on whom the health effects of air pollution were followed prospectively, beginning in 1974. The objective of the study was to estimate the effects of air pollution on mortality when controlling for age, sex, individual smoking status, and other risk factors. The study found that mortality risk was strongly associated with the levels of fine particulate air pollution (particles smaller than 2.5 microns, or PM2.5). This research has led to EPA air quality standards and regulations on PM2.5.

they are because it intersects with the air quality data at that time when they might be admitted." The combination of the various types of data made the study particularly valuable for policy makers, Casey said, because it allowed researchers to draw connections between a person's exposure to air pollution and that person's health and (for those who died before the study was finished) age at death.

But the combination of information also made it much easier to identify the subjects in the data set, he noted, "because if you know what the air quality is at a particular time and you know when somebody died in a particular day and you know the town where they were located in, it is not very hard.... Two people died in Watertown on July 2, 1997. One is a woman. One is a man. We actually know who it is. We provide [those] data. It violates exactly what we had promised when these people enrolled."

Casey added that the death index data were also provided with a promise from the researchers that they would not allow the data to be used in a way that would disclose any individuals who were part of that work.

But the situation sets up a tension between the importance of using the data for scientific and regulatory purposes and the need for maintaining the privacy of individuals who provided their data with the promise that their identities would not be revealed. The data from the study are "foundational information" used in setting clean air rules, Casey said. "And the Clean Air Act states that it should be revised periodically, and, when it is, it should draw upon the published research that is available, and when they do, it has directed attention to the Harvard School of Public Health. Our work percolates. It bubbles back up." Unfortunately, the need to protect the confidentiality of the subjects enrolled in the study leads to what Casey described as "our inability to provide enough information to certain policy makers to make them feel like they have all of the data to support the determinations of EPA and others on clean air standards." This in turn creates pressure to provide more data, which could make breaches of confidentiality more likely.

"It is a major issue," Casey said, "because the whole concept of enrolling people in studies is at risk." People are likely to respond much differently if they are told that their data will be kept private than if they are told that their data might be provided to members of Congress and other people who will use it as they like. "I think that it would have a chilling impact," he said. "It is not to cast aspersions on the Congress or others. It is an actual conundrum that we are facing."

The issues being raised by members of Congress are legitimate, Casey said. It is true that there are certain data that the researchers have not been willing to supply. "The reasonable-thinking people who are serious about clean air, who are serious about solid policy, who want the federal government's work to be valid—those people have reasonable questions that they are raising," he said. "We would like to get to the bottom of them, also. Our researchers are not trying to hide the ball. Some of them have been working for 20 years and have not gone to a meeting where they have not been confronted with this issue. If there were a way that they could thread that needle and provide the amount of information to satisfy reasonable policy makers in the pledge that they made to their subjects, they would do it tomorrow."

COMMENTS AND DISCUSSION

This section summarizes the discussions on the challenges associated with data sharing that took place throughout the workshop.

Issues Associated with Reanalysis of Data

A workshop participant watching the workshop over the Web noted during the discussion after Session 2 of the workshop that one of the concerns that researchers have about sharing their data is that whoever does a reanalysis of the data may not do it to particularly high standards and come up with results that are incorrect. This is a particular worry when advocates for a certain side of an issue are looking at the data, but the problem is broader than that and extends to researchers who may not have a particular axe to grind but are simply not being very careful in their analyses. Specifically, how does one ensure the quality of the reanalysis of data? The example offered was a reanalysis of the association between vaccines and autism in children that found that an association did exist. Although the reanalysis was done poorly, its results were quickly picked up by the press and may have contributed to mothers choosing not to vaccinate their children. Gwen Collman of NIEHS suggested that any reanalysis should be subjected to the same standards as the original analysis. They should go through peer review before they are released and certainly before they are used in any sort of rule making by either the regulatory agencies or the courts. Anthony Cox, chief science officer of NextHealth Technologies, noted that the original analyses are often not of very high quality, which is why the

scientific literature has so many false positives. Logically, he added, false negatives and false positives are equally important, but, empirically, false positives are the problem, which is why he focused on them. Finally, he commented that he agreed with Collman that "all analyses that are published should have to jump through the same hoops."

Al McGartland of EPA commented that there is a similar problem with reports in the gray literature, that is, literature containing the findings of research that is reported by associations or companies but that has not gone through peer review. Although these studies are not published in academic journals, they are often circulated on Capitol Hill and cited by the press. Should the researchers who publish gray literature release their data and models for use in reanalyses? Cox responded by stating, "I do not think there is a good, clean answer other than buyer beware or reader beware." If the peer-reviewed scientific literature can be cleaned up and made "closer to the gold standard that we all want it to be," Cox said, then the peer-reviewed literature will inevitably have increasing value compared to the gray literature and the gray literature will get less attention.

Informed Consent

During the discussion after Session 2 of the workshop, Alan Morrison of the George Washington University Law School highlighted two basic issues with informed consent as it concerns environmental health data. "The real problem today," he said, "is that when you ask somebody for consent, particularly a broad consent, neither you, the requestor, nor the person being requested has any notion at all as to what that means in terms of how it is going to be used because we do not even know what it is going to be used for down the road. The second thing is [that] when people asked for consent a long time ago, we understood that consent meant you were not going to hand out something to somebody else, but as long as we took your name off it, [it] was okay. Obviously, that is not enough anymore." With all of the ways that people in data sets can be identified that have been developed, even if the identifying data have been removed, people who were once essentially unidentifiable are now identifiable. This changes the game and means that the consent that people gave 20 years ago can no longer be considered "informed" consent because neither the subjects nor the researchers had any way of knowing just how identifiable the records in data sets would become. Thus, Morrison concluded, "the notion of informed consent as it was

developed in the law is, in my view, essentially meaningless. We have to think about it in a different way."

Bernard Lo, president and chief executive officer of The Greenwall Foundation, agreed with Morrison on both issues. Concerning the informed consent that researchers today are obtaining from subjects for the sharing of their data, he said that there is a tendency to promise too much. "By implying that we will only do what you consent to and that we will protect your privacy in very strict terms so there is no probability that people can reidentify you may be promising too much. It is unrealistic."

As for the informed consent provided by subjects in older research studies, Lo agreed that it is questionable whether the consent can hold up today. "We heard data presented from investigators saying they looked retrospectively at consent forms for completed drug clinical trials. What did they say about sharing? It was split. Some trials did not mention it at all. They were silent.... Others gave broad consent: 'This could be shared with other researchers.' The participants may not have known what that means." So what should be done with clinical trials where the original consent did not mention data sharing or said, for example, that the data should be shared only with other researchers in the same institution? It is not clear.

Linda Birnbaum, director of NIEHS of NIH, noted that the U.S. Department of Health and Human Services is currently working on the issue of informed consent and, in particular, is working on a proposed new rule related to changes in the informed-consent language that would offer broader informed consent, with researchers being given the opportunity to go back to the participants to ask them for permission to use the data in a new way each time that the researchers wish to reanalyze the data.

In the opening remarks for Session 3 of the workshop, Glenn Paulson, science adviser in the Office of the Administrator at EPA, said that there is a new working group on modernizing the Common Rule, the federal rule governing the treatment of research subjects in many federal departments. Changes in the Common Rule could affect such issues as informed consent and the sharing of data from surveys of trials involving human subjects.

The Harms of Reidentification

Assuming that some records in a data set are reidentified, the obvious question to ask is: What are the potential harms associated with that reidentification? What damage can be done to people who contributed data to a data set if their identities become known?

John Gardenier, a retired member of the confidentiality staff at the National Center for Health Statistics, raised the issue in a question that he posed to Daniel Barth-Jones in the discussion after Session 3 of the workshop: "We can all hypothesize the theoretical or hypothetical harms that could come from a false-positive reidentification," he said, "but I do not know of any specific study that has been done in which anybody has documented an actual case of actual harm as an example of what the real societal risk is. Can you comment on that?"

Barth-Jones answered that he also was unaware of a demonstration of actual harm that reidentification had caused to a real person. "I think our evidence base at this point is fairly slim," he said. "One of the things I would argue for is doing more systematic studies where we actually do good sampling and then look at how many people can be reidentified. And then I think we are at the point of having to make some subjective assessments of the sensitivity of the information that may be revealed and whether it could lead to potential harms."

This discussion led to a series of other comments concerning the harms of reidentification. One workshop participant suggested that harms can come about in ways other than people in a data set being identified. "I think another model that exists out there is, if I know an individual, can I go to this data set and find out if that individual is in the data set?" This might come up, for example, if a plaintiff in a lawsuit claiming harms from environmental exposures of concern had taken part in an environmental health study. The opposing lawyer might search through the data set from the study, looking for additional information on the plaintiff, which would be accessible if it was possible to identify the plaintiff from among the people who took part in the study.

John Howard, director of the National Institute for Occupational Safety and Health, offered a second example. An employer could search for his employees or potential employees in a data set from an occupational health study in an effort to get information about the disease status of those people.

Howard also suggested that there does not have to be this sort of real-world harm for harm to have happened. Once a person who has been

promised privacy is identified within a data set, that breach of the promise of confidentiality is itself a harm, he said.

Barth-Jones agreed and described a similar situation that arose with reidentification of people taking part in the Personal Genome Project. "Even though there was no promise of confidentiality in that situation," Barth-Jones said, "many of the people who were reidentified in that situation without having been consulted first seemed to express displeasure."

Reidentification Risk as a Participation Deterrent

The individuals whose data are in a data set are not the only ones at risk from reidentification. If reidentification becomes a serious issue, it is possible that this could make it less likely that individuals will be willing to take part in studies.

Stacy-Ann Allen-Ramdial, intern with the House Committee on Science, Space, and Technology, raised the question explicitly during the discussion after Session 3 of the workshop: "I have heard a lot about the importance of sharing data for researchers and for the federal agency," she said, "but wouldn't the participants be a little bit more comfortable with the transparency early on, knowing that their information could potentially be used by the federal government for regulation? Or is there a risk that you will lose so many participants that you will not have enough for your research studies?"

Casey answered that people are trying to find the proper balance between the need for to share data and privacy concerns, but he suggested that it would have a "chilling effect" on people considering enrolling in a study that would take their personal information and study them over many years "if they knew that their personal information might become part of the public discourse record—individual information, not aggregated, not collected and shared in terms of overall studies." Researchers have not really had to deal with this issue up to now, he said, and so there is no direct experience that indicates what would happen. However, he added, "I suspect that that would have an actual damaging impact on the ability to enroll people, not only in this kind of [clean air] study but potentially downstream in other kinds of clinical trials and other areas where things could become controversial."

This exploitation should have its limits, though, warned Casey. For example, the data in the Harvard Six Cities Study have been, if anything, overanalyzed in the two decades since the study appeared. "The original

publication was 1993," he said, "and there continues to be a fetish about reanalyzing that initial base of data when there have been over 20 studies done over the last 20 years by international and national groups that have come to similar conclusions and have drawn upon different sets of participants in the groups. To continually go back to the 1993 data is acting as if science was frozen two decades ago."

Glenn Paulson, science adviser at EPA's Office of the Administrator, offered his own experience from serving on the executive committee for an institutional review board at a major medical research institution. The institution carried out a wide range of studies, from clinical trials to epidemiological and behavioral research. Recruiting people to participate in research projects was always a problem, he said, but it never seemed to be due to the worries about privacy and confidentiality issues. The consent forms generally said that information about the subjects would be kept private "unless it is requested by a court of law or something to that effect," he said, and this language did not seem to be a reason for people not being willing to join research studies.

Lynn Goldman, dean of the Milken Institute School of Public Health at George Washington University, spoke of the repercussions of a ruling to publicize data that a researcher had promised to keep private. In 2005, Bruce Lanphear and colleagues published a study looking at the effects of lead on intelligence quotient (IQ) in children. Working with colleagues who had collected data from children around the world, he pooled the data and found that exposure to lead lowered IQ in children (Lanphear et al., 2005). Because EPA had funded Lanphear's systematic review, the decision was made that the data in his study had to be shared, so without the consent of any of his colleagues who had provided the data—and without the consent of the subjects who had provided the data—the data were made public to anybody who wanted to use them. "My guess is that he [Lanphear] repaired the relationships with those other investigators, and that they are okay," Goldman said, "but there are people around the world who may feel cautious about collaborating with U.S. scientists in any setting where they might be asked to do a data-sharing agreement, knowing that if the federal government is funding it, their data then may become just available freely."

"I do not think that that is helpful in terms of generating good will about scientific collaboration and cooperation among countries," Goldman continued, "or for that matter eagerness for participating in systematic review. I feel that is a harm, to be honest."

Ellen Silbergeld, professor at the Johns Hopkins Bloomberg School of Public Health and editor-in-chief of *Environmental Research*, added that, for the sorts of studies that she does, this situation would make it very difficult to get people to volunteer for studies. "If I told anybody in the studies we do with Native Americans, people in the Amazon, people in inner city Baltimore, and workers in South Carolina that the data might become available through a lawsuit, they would not participate."

During her presentation in Session 5 of the workshop, Birnbaum suggested that the chilling effect is most likely to be seen in smaller studies involving people who are followed for many years. "This may not be as much of a problem in some of the very large, statistically based, almost ecological kinds of epidemiological studies," she said, "but it may be extremely important when we are dealing with the smaller studies where we have small cohorts that we recruit and we want to follow."

Explaining Risks Better to Subjects

Howard commented during the discussion after Session 3 of the workshop that it is no longer possible to believe—as it was several decades ago—that the confidentiality of research subjects can be absolutely assured and that researchers thus have an obligation to talk about confidentiality risks differently than they did many years ago. "It would seem to me," he said, "that we are at a stage of developing obligation on the part of scientists [where] they should know that the situation of making those kinds of promises to a potential study participant in 2014 versus 1970 really isn't the reality of the world we live in." In particular, he said, it is important to educate researchers on the best way to talk to subjects about confidentiality risks, but it is not clear to him that many institutions are making sure that their researchers are up to speed. The situation concerning confidentiality risks has changed so much in the past 10 to 15 years that it may be the case that many scientists are still making promises to their subjects that they may not be able to keep.

REFERENCES

Barth-Jones, D. C. 2014. *Challenges associated with data-sharing: HIPAA deidentification*. Presentation at the workshop Principles and Obstacles for Sharing Data from Environmental Health Research, Washington, DC.

Dockery, D. W., C. A. Pope, X. Xu, J. D. Spengler, J. H. Ware, M. E. Fay, B. G. Ferris, Jr., and F. E. Speizer. 1993. An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine* 329(24):1753–1759.

El Emam, K., E. Jonker, L. Arbuckle, and B. Malin. 2011. A systematic review of re-identification attacks on health data. *PLoS ONE* 6(12):e28071. Available at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0028071 (accessed October 27, 2015).

Lanphear, B., R. Hornung, J. Khoury, K. Yolton, P. Baghurst, D. C. Bellinger, R. L. Canfield, K. N. Dietrich, R. Bornschein, T. Greene, S. J. Rothenberg, H. L. Needleman, L. Schnaas, G. Wasserman, J. Graziano, and R. Roberts. 2005. Low-level environmental lead exposure and children's intellectual function: An international pooled analysis. *Environmental Health Perspectives* 113(7):894–899. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1257652 (accessed October 27, 2015).

Sweeney, L. 2014. *Inconvenient truths of re-identification discourse*. Presentation at the workshop Principles and Best Obstacles for Sharing Data from Environmental Health Research, Washington, DC.

5

Possible Ways Forward

This chapter focuses on the discussions from Sessions 4 and 5 of the workshop, where participants offered their thoughts on how the field of environmental health might best move forward in the sharing of data and reflections on the workshop. The final section of this chapter, "The Bigger Picture," is a summary of the remarks that one presenter provided during Session 5.

WAYS TO INCREASE DATA SHARING

Several attendees provided some principles and lessons that should be kept in mind when thinking about the topic of sharing environmental health data. Unless otherwise noted, all comments summarized in this section were made during the discussion after Session 4 of the workshop.

Ensuring Quality Data Sharing Practices

John Howard, director of the National Institute for Occupational Safety and Health (NIOSH), stated that there should be a balance between reanalyzing the old, existing data and looking at new data. "From the science perspective, we want people to continue to look at data with new data in mind and come to different conclusions so that it is a living organic piece of knowledge," he said during the discussion after Session 3. "That is why we promote and are trying to figure out a way to enable reanalysis. Hopefully, studies will not be looking at the data from 20 years ago without taking into consideration new data sets that have enriched the original data set."

Lynn Goldman, dean of the Milken Institute School of Public Health at George Washington University, noted that when raw data are released, they are generally not completely raw. Instead, they have been cleaned up to a certain degree. "Data cleaning is ... very difficult to do even when

it is your own work," she said, so few researchers would want to have to clean other researchers' data. "I am pretty sure National Center for Health Statistics does not release data that [have] not been cleaned," she said. "The quality control is done before it is made available to the public. But I do not think that that has been clearly articulated."

George Daston, Victor Mills Society Research Fellow at Procter & Gamble, expanded on that by noting the importance of quality control in both the collection of the data and the processing of the data. A major factor in whether researchers and others believe in data that have been collected, he said, is how much faith they have in "what sorts of quality control have been applied to both the data collection and the processing." Thus, it would make sense to think about standardizing such quality control.

Francesca Dominici, a professor of biostatistics and senior associate dean for research at the Harvard University School of Public Health, agreed with Daston that the field of environmental health needs these tremendous quantities of data if it is to keep advancing. "In environmental health, we do not have low-hanging fruit anymore," she said, mentioning the early work on the connection between smoking and lung cancer as an example of such low-hanging fruit, since the effect is so large that it is relatively easy to find. Now, she said, environmental health scientists are trying to assess whether or not various environmental contaminants are still harmful to human health at very low concentrations. "To do that," she said, "it requires terabytes and terabytes of very complex data. I think everybody will agree with that."

Daston replied that he agreed with her comment that the low-hanging fruit is gone and that the field is past the time of finding either one gene or one environmental agent that causes a disease. "If we really want to make headway in public health," he said, "it is going to be through sharing large data sets."

The way in which the data are collected is a crucial part of data sharing, said Edward Sondik, former director of the National Center for Health Statistics at the Centers for Disease Control and Prevention. He described hearing a debate about some Canadian breast cancer data that had recently come out. "The issue between somebody who felt very strongly that the data were flawed and the Canadian investigator really had nothing to do with the data per se," he said. "It had to do with the way the data were actually collected. It went back to ... the actual structure of the study." Thus, it is crucial that a study's metadata be shared because they contain information about how the data were

collected, that is, about the structure of the study. "If you agree on the structure, then you are dealing over the analytic methods," Sondik said. "But in many cases that I have seen, the issue is not the methods per se. It really is how the data were actually collected."

Ellen Silbergeld, a professor at the Johns Hopkins Bloomberg School of Public Health and editor-in-chief of *Environmental Research*, argued that while the sharing of data for the purpose of reanalysis may be necessary at times to increase trust in scientific results and the regulations that follow from them, reanalysis of data is not a particularly useful part of science. "I do not think science is advanced by reanalyzing data sets," she said. "Science is advanced by people replicating studies in different situations with different populations that are completely independent."

"I think we are kidding ourselves," she continued. "This has very little to do with the quality of science. I do not think anything was advanced by extracting all the data from the Needleman studies,[1] reanalyzing them 15 times, accusing him [Needleman] of fraud, and then moving on. What was advanced by having 20 more studies?" Science is not advanced when researchers go back to the Harvard Six Cities Study and look at their data again, she said. Instead, science advances through the accumulation of better analytic methods and statistical methods that allow one to provide a better answer to the same question that was being asked. "Could we stop saying that this is a matter of advancing science? ... I would challenge anybody to show me where an analysis of existing data was as earth shaking as doing a really fantastic study that looked at the same hypothesis and either replicated it or moved it forward into a different area."

[1] Herbert Needleman and colleagues conducted a study in 1979 that investigated the effects of lead exposure and toxicity in children. Children with elevated lead levels were found to be significantly impaired on intelligence quotient (IQ) tests and exhibited negative classroom behavior (Needleman et al., 1979). In the 1980s, they employed larger samples and more sophisticated statistical analyses to continue studying the issue, and three meta-analyses found that low-level lead exposure was associated with IQ deficits (Needleman and Gatsonis, 1990; Pocock et al., 1994; Schwartz, 1993). These studies played a critical role in the elimination of lead from gasoline and the lowering of the Centers for Disease Control and Prevention acceptable blood lead standard in children.

Developing Common Language and Standards

To move forward effectively on sharing environmental health data, it will be helpful if everyone is speaking the same language and adhering to the same standards, said Linda Birnbaum, director of the National Institute of Environmental Health Sciences of the National Institutes of Health (NIH), during her presentation in Session 4 of the workshop.

"This is a huge issue," she said. "We actually held a data ontology workshop a couple of months ago to begin to develop a common understanding of what we mean.... What do you mean by data? What do you mean by reanalysis or replication or reproducible? I think that a common language for environmental health would foster the interoperability of databases and promote the sharing, reuse, and reanalysis of data and therefore, hopefully, accelerate the pace of discovery."

Several workshop participants commented that the definitions that Goldman had offered earlier for "reanalysis," "replication," and "reproduction" (see Chapter 2) were good ones and should be promulgated. "You were very close in your definition," said Daniel Greenbaum, president of the Health Effects Institute. "I think it is important, and I do hope that the Roundtable ... could try and clarify that and make it pretty important." In response, Goldman clarified that the definitions were from the Oxford English Dictionary and were not hers. Greenbaum did suggest a slight modification for one of the definitions: "In order to reanalyze, i.e., to do all kinds of sensitivity analyses and other things, not just check the math, the first thing you have to do is replicate the original results," he said. "You have to ask: Can I do exactly what they did in this data set and come up with the same thing so I know I am working on the same data set? ... Reanalysis goes much further than just checking their math."

On a related topic, a few workshop participants suggested that the development of standards for how data are submitted and represented could make it much easier for multiple researchers to work with the same data. "Should we begin to work with the publishing community about ... standards for submitting studies for journal publication?" Birnbaum asked during her presentation in Session 4 of the workshop. "Could there be standardized formatting—for example, for the methods and the key findings—that would help in reporting quality and make automatic curation more feasible?" Plenty of available text-mining software makes it possible to search thousands of different texts for information on a particular subject, and attaching a list of standardized key words to journal articles—or using the standardized key words inside the

articles—could greatly improve the efficiency and accuracy of the text-mining software. "This could allow for the automated curation of published findings into the databases," she said, "and when you have that, that could become a research tool that could be used again."

As part of those standards for data, some workshop participants suggested that it would make sense for a study's metadata to include information about conflicts of interest among the study's authors. "For every single paper that is published in a peer-review journal, you can ask how much the investigator put in his pocket by producing this study," Dominici said. "I think most of the academic world would have no problem in disclosing how much money they make when writing a single paper." Birnbaum suggested that each author of a study should report how much he or she was paid for carrying out and writing up the research and where those funds came from—whether from a research grant; salary from a university, company, or government agency; consulting fees from some advocacy organization; or somewhere else. "I think it is absolutely essential that we begin to have that information fully available," she said. "If anyone says that who puts the food on your table does not have some kind of impact on your gestalt, I am going to find that hard to believe."

Planning and Time Limits for Data Availability

Birnbaum pointed out that it is important to think about *when* data should be made available for others to work with. "I know our investigators have huge concerns about premature release of data or premature demand for release of data before [they are] fully analyzed," she said, adding that by "fully analyzed" she did not mean that the investigators had extracted every implication from the data that they could but, rather, that the investigators examined the data thoroughly enough to be comfortable that they were reasonably error free and ready for analysis by other parties. But while it is important to make sure that the data are cleaned up and vetted before they are released, it is also important that the cleaning-up process not unreasonably delay the release of the data. It is a balance that needs to be thought about, Birnbaum said.

Greenbaum suggested that a basic principle should be that "early on in the process, right as you go to fund a study, [you should be] expecting to build into the grant and into the proposal the plans for making data available at the other end and the dollars that are going to be necessary to do that." That is not always done today, he said, but it should be done by

anyone who is funding research in either the private sector or the public sector.

Even when data are made available to other researchers, there are limits on how long the data will be or can be shared. "We are almost going to have to declare some form of statute of limitations on how far back we can reach in reaching old data," Greenbaum said.

There are two issues, he said. The first is the practical issue of being able to read data that were stored in certain formats—say, on floppy disks or magnetic tapes—many years ago. At a certain point it becomes so difficult and expensive to resurrect the ability to read data in such formats that it is simply not feasible. Thus, unless early data sets have been transferred into more recent forms of storage, researchers might not be able to access them.

The second issue is that federal policies do not require data from scientific studies to be stored for a particularly long time. Greenbaum said that he believed that it is generally only 7 years. "I am not talking just about internal data retention policies," he said. "I am talking about grantees getting money." Thus, data that the federal government paid for may not be available for researchers who come along a decade later and need them for some new analysis. Thus, it will be important, Greenbaum said, to develop principles for dealing with such situations when the results of the studies are available but the data underlying them are not. "We are just not going to get some of the data," he said. "It does not mean the study is invalid or cannot be used. That is an important thing."

Given this situation, he said, it will also be important going forward to think about how long data should be made available.

Reducing Tensions Associated with Data Sharing

Greenbaum discussed the roots of the tensions between researchers and some policy advocates and possible ways to approach lessening of those tensions. "The reason there is so much attention to certain studies is usually because there are incredible policy and economic stakes involved in decisions that are based on those studies," he said. In a democratic and often adversarial society, there should be ways for people who have concerns about government decisions to raise those concerns. Indeed, our system of government is structured to allow individuals and organizations with concerns to have their voices heard. "EPA's [the U.S. Environmental Protection Agency's] rules get challenged all the time on a variety of factors, not just on the health basis," he said.

The challenges to agency decisions are often not based on whether the underlying science was done correctly, Greenbaum said, but rather on whether the science was considered in a thoughtful and rational manner by the agency making the decision. "Sometimes EPA loses, and sometimes they do not. Sometimes OSHA [the Occupational Safety and Health Administration] loses, and sometimes they do not."

While acknowledging that he did not know whether there would be any way of easing tensions, Greenbaum suggested that one way of easing tensions would be to "construct rules of engagement that promoted a level of civil discourse that enabled people to actually produce quality science, have it challenged by scientists no matter who they worked for, but in a scientific manner, have dialogue and opportunity for dialogue, and then in the end know something more than we did before as a result of that." Greenbaum said that the reanalyses of the Six Cities Study data that his group did had some of those characteristics, so he recognized the value of such scientifically based challenges, but, at the same time, he also recognized that it is not feasible to carry out such reanalyses for every single study that a government agency supports. "There just aren't the resources to do that."

Greenbaum also suggested that the National Academies of Sciences, Engineering, and Medicine is the right kind of place to hold a discussion on how to develop such rules of engagement. Any such discussion would require a core group of people who understand what is acceptable and what is not acceptable practice, he said. For example, bringing a scientific challenge against a legitimate investigator just to intimidate that investigator crosses the line, he said. "On the other hand, asking to see the underlying data to do reanalyses of it to understand what it is and to go through getting it published in the peer-review literature, even if you are advancing one of the adversarial sides, has the potential to be very positive if it is done with the right rules of engagement."

Greenbaum added that he would not trust the current Congress to determine what those rules would be, emphasizing that he was not talking about one party or the other. It would be better to have groups of scientists carry out such discussions and propose appropriate rules. "There was a group [organized by the Bipartisan Policy Center] that Lynn [Goldman] and I were on a few years ago that wrote a report on advancing science for policy purposes that tried to lay out some of the principles," he noted. "There may be others."

Ultimately, Greenbaum said, science at its best is a highly adversarial undertaking "not for the reasons we are talking about, but

because different scientists have different views of what is the right answer and what is the wrong answer." The goal should be to move the country's political adversarial process in a direction where it would be more like a scientific adversarial process. "We have rabid debates among scientists over whether somebody's study is correct or not correct, and that goes on all the time," he said. Could the disagreements over scientifically based policies be carried out in a similar manner? It is worth finding out.

Providing Incentives to Share Data

Hal Zenick of EPA stated the need to develop metrics to measure data sharing. "If you look at the baby boomer cohort," he said, data sharing "is a foreign concept. During our careers it was a very competitive type of atmosphere. Your publishing was your metric. Publications, presentations, funding were all based on competing. If we fast forward about 20 to 30 years, we have a whole new cohort. That cohort lives social media. All they do is data share back and forth. But the incentives for that have not emerged." Most of the metrics used today to assess researchers—publications, funding, and number of presentations—are rooted in the era of competition, he said, and there are few metrics to measure data sharing or the results of data sharing. It would be useful, Zenick said, to begin having discussions about the best way to measure data sharing by researchers and what that data sharing led to. That will be a necessary first step toward creating real incentives for researchers to share their data.

Another sort of incentive to sharing data—which would be particularly aimed at policy makers—would be to objectively demonstrate the benefits of data sharing, said Daston. "There aren't any 100 percent positive aspects of data sharing," he said, "but what we have to do is get to the point where we can do the calculus that shows that the benefits of the data sharing far outweigh whatever the costs are." Furthermore, he suggested, it would make sense to pay attention to who bears the costs of data sharing and "if the costs are disproportionately borne by one individual or one sector, find ways to compensate them." But it will be clear demonstrations of the benefits of data sharing that will carry the most weight in convincing people to push for more.

Utilizing Secure Enclaves

Goldman suggested that since there are already data enclaves for data collected by federal agencies, it would make sense to establish similar enclaves for data from investigators who are funded by the federal government. "Is it possible," she asked, "that that kind of an enclave could be a home where, when you submit your paper, you could submit your data and then you wouldn't have to worry about looking after this generation of technology 30 years from now? Somebody else would be responsible for keeping it as long as it ought to be kept, retaining it, keeping it up to date, keeping it available. And might that perhaps put that process at arm's length between whoever is trying to access the data and the investigator?" If feasible, it would be worth looking into.

Greenbaum replied that it was an interesting idea. "One of the issues that that raises is that if a federal agency itself has conducted a study, constructed the data sets, et cetera, and then chooses to share its data either through a data set or some other mechanism, there are a series of laws that say people have to sign data user agreements" to have access to the data. There are also rules for NIH data sharing more broadly, he said. "I do not know that that is so clear-cut for a federally funded study at Harvard or Stanford or for somebody else who has a data set. When I signed an agreement with Harvard, I did not have a federal law" that specified what the data user could and could not do with the data.

Thus, Greenbaum suggested that if the federal government provided data enclaves for the sharing of data from federally funded studies, it could lay out the same sort of user agreements, and if a user broke the agreement and breached the confidentiality of people whose data were in the data set, the same sort of federal penalties, which are up to $250,000 or 5 years in prison, or both, could apply. "That would help protect privacy."

OVERALL REFLECTIONS

During the Session 5 discussion, workshop attendees provided reflections on the workshop as a whole.

Take Advantage of What Is Already Out There

Joseph Rodricks, principal of ENVIRON, pointed out that a number of groups, both in environmental health and in other areas, have spent time grappling with issues surrounding the sharing of data. "One of my colleagues ... pointed out a very interesting document from the Oak Ridge National Laboratory on best practices for preparing environmental data sets to share and archive,"[2] he said. "It is a big document that goes into exquisite detail on what investigators should be doing to prepare to publicly release the data as they develop [them]. It is excellent guidance, I think." It would make sense to pay attention to these sorts of documents to avoid reinventing the wheel, he suggested.

On the technical side, Jerry Blancato, director of the Office of Science and Information Management at EPA's Office of Research and Development, noted that a variety of technical approaches to sharing data without breaching confidentiality have already been developed. "There may be ways that data can be shared publicly, but you still protect the pieces of data," he said. "I do not think it is our business as researchers to reinvent the wheel. There are experts out there, industry experts who not only have the capacity to store tremendous amounts of data and do it reasonably cheaply but have the wherewithal to do it.... We have to be able to take advantage with those partnerships to get the data there."

Effects of the Coming Data Tsunami

Latanya Sweeney, professor of government and technology in residence at Harvard University, suggested that the rapidly growing availability of personal data from a large number of sources is changing the equation about personal privacy and confidentiality and said that much of the discussion that she heard at the workshop seemed outdated because it was not taking these changes into account. One example, she said, is the Personal Genome Project, the goal of which is to collect and make public genome data and the detailed medical records of 100,000 volunteers to accelerate research into personalized medicine and personal genomes. Another example is the Fitbit, a device that keeps track of a wealth of activity on the people who choose to wear it, such as daily activity patterns and levels, calories burned, sleep patterns, and weight.

[2] "Best Practices for Preparing Environmental Data Sets to Share and Archive" is available at http://daac.ornl.gov/PI/BestPractices-2010.pdf (accessed February 22, 2016).

While taking part in the Personal Genome Project and wearing a Fitbit are personal choices, the world is moving toward gathering data about people in all sorts of ways, many of which they will not be aware of, said Sweeney. "This will definitely change the way you will conduct research and grab data, who you will get the data from, and also privacy and other kinds of issues." Realistic discussions about privacy and confidentiality will have to take into account the coming changes in the ways in which data are collected, the types of data that are collected, and the attitudes that people have about their data being collected.

Organizing the Scientific Community to Take Action

Frank Loy noted that the scientific community should be proactive in the debates between Congress and EPA around data sharing to achieve a reasonable balance between transparency and protection of data. "The scientific, medical, and university communities have status in this country that, quite frankly, EPA does not. Therefore, I would like to see this treated as a science issue rather than as an EPA regulation issue."

Goldman agreed with this point and noted that as a scientific community, "We have choices about what we can do.... We have been doing nothing actually to affirmatively address this issue. We have been placing ourselves in a position where therefore Congress or others are trying to impose things to take care of this." She noted that it can be particularly challenging to engage the environmental health community in issues like this. "Quite frankly, we [environmental health researchers] are not a community. We are in many different professions, different professional organizations, and are not exactly organized in any way, shape, or fashion in the way that AAAS [the American Association for the Advancement of Science] and other large science organizations are organized." Goldman stated that rather than being passive and watching these events occur, it would be better if there was a way to organize the environmental health community around data sharing. "I would like for us all to be thinking about ideas that could be moved forward to bring people together to try to address this. In other words, can we be in charge of our own destiny?"

THE BIGGER PICTURE

During her presentation in Session 5, Silbergeld offered a cautionary take on what she saw as basic underlying assumptions at the workshop.

To begin, she noted that this workshop topic has potential costs to the scientific community, the excellence of science, and the creativity of science in the United States and is something that should be given much more thought than it has been given to date. She found it extraordinary that at the beginning of the workshop nobody defined "what do you mean by 'data,'" making it difficult to understand what is going to be shared. Before moving forward, it would be good to define the term with some specificity or at least give a range of what it could mean, she said.

Silbergeld offered some sobering comments that indicated what is really at stake if more thought is not given to this topic. "I have to say I come away from this meeting thinking that I probably will never do another environmental epidemiological study," she said. "I cannot stand behind the guarantee that I have always felt to be important for research integrity [that I am able] to protect the confidentiality of the people who are brave and caring and patient enough to get involved in the studies that we do in environmental epidemiology." Furthermore, she suggested, U.S. researchers may find it difficult to attract international collaborators if confidentiality cannot be ensured. "I will think twice about engaging my colleagues in other countries in joint ventures in which their data could become accessible to adversarial proceedings in the United States," she said.

There is a "bigger picture" that should be remembered. "It seems to me," Silbergeld said, "that in my career of being involved in environmental sciences, it has always been about ensuring the highest-quality data [are] going into making of decisions that have both economic and public health impact. I am not sure that ensuring excessive and complete access to data sets is the way forward to reach that goal most consistently and expeditiously. I think there are paths forward that we could think about and certainly ones where we can learn from other disciplines."

In particular, she suggested, the model that should be kept in mind is evidence-based medicine. "This has been an experiment of some 60 years in the making ... [and] was also meant to deal with a highly contentious topic, which was the advent of national health care in the United Kingdom." As the United Kingdom was instituting its national health care program, an economist, Sir Stafford Cripps, posed the question of how one should decide which treatments should be paid for. Different doctors had different answers, and none were able to offer any objective evidence supporting their answers. Cripps set out to find a

better way, and "that was really the birth of evidence-based medicine," Silbergeld said.

The development of evidence-based method led to a particular methodology that can be applied to environmental health research, she noted. First and foremost, Silbergeld said, that methodology is transparent. "I understand we have a great deal of transparency," she said, referring to people in the environmental health field. "Although sometimes we have different places where transparency begins and ends, ... it is very much based on the insights of several stakeholders. One is the scientists and the generators of knowledge, and the others are the people who use the knowledge (for example, in medicine) and the patient community (as well as their values and concerns)." Silbergeld stated, "[this] is where I come back to this issue about the respect that we owe our subjects in environmental epidemiology—a respect that is in danger of being taken away."

Silbergeld then discussed how one should go about carrying out a systematic review in environmental health. "The goal of a systematic review is to get as much information as possible—not just set up barriers towards the admission of information but at the outset to have a fair amount of confidence that we have surveyed the entire body of available information," she said. Researchers will establish certain boundaries, of course, "but within those definitional boundaries, we have a pretty good degree of confidence that we know what that landscape looks like." To that landscape the researcher applies a set of predetermined analysis criteria that are fixed ahead of time to the information that has been developed. "We do not make it up as we go along," she said.

"Now I ask my question: What do you mean by data, and what do you mean by information? The whole goal of the systematic review is to translate information into evidence," with "evidence" being "information in which we have confidence," Silbergeld said. Such confidence arises by reducing bias to the greatest extent possible, by understanding the strengths and the limitations of the information, and by applying certain criteria, such as those set forth for good laboratory practices or the most extensive sets of criteria that have been developed for both clinical trials and observational epidemiology. "We can then use those in an extremely transparent way to say, 'This is the information that we are going to consider converting into evidence.'"

Sometimes the process requires going to the researchers who have published the papers in question and asking them for some additional information to help clarify their use of their particular bits of information.

"We do that with a great deal of trust among the people who have generated that information and those of us who want to use it," she said. Silbergeld continued, "And from there, we can move forward to something that I think would improve not only decision making but also gives the yardstick as to how much information you need. I do not think you need the raw data tables." There may be particular cases in which there are reasons that the raw data are needed, she said, "but in the general evaluation of both toxicologic and epidemiologic data, I do not think we need them."

Silbergeld stated that the larger goal of maintaining a steady supply of data through new innovative studies should be kept in mind when individuals choose to gather data from researchers. She worries that some of the actions that she heard talked about in the workshop could make it less likely that the supply of important new data will continue unabated. "To suggest that scientists should start a study figuring out how they are going to make all their data available at the end will have a very interesting effect on how we train the next generation of scientists. I can tell you that. I think there is another path forward," she said.

REFERENCES

Needleman, H. L., and C. Gatsonis. 1990. Low level lead exposure and the IQ of children. *Journal of the American Medical Association* 263(5):673–678.

Needleman, H. L., C. Gunnoe, A. Leviton, R. Reed, H. Peresie, C. Maher, and P. Barrett. 1979. Deficits in psychological and class-room performance of children with elevated dentine lead levels. *New England Journal of Medicine* 300:689–695.

Pocock, S. J., M. Smith, and P. Baghurst. 1994. Environmental lead and children's intelligence: A systematic review of the epidemiological evidence. *BMJ* 309:11889–11897.

Schwartz, J. 1993. Beyond LOEL's, p values and vote counting: Methods for looking at the shapes and strengths of associations. *Neurotoxicology* 14:237–246.

A

Agenda

March 19, 2014
Lecture Room
National Academy of Sciences Building
2100 Constitution Avenue, NW
Washington, DC

8:30 a.m.	**Welcome** Frank Loy, LL.B. Roundtable Chair
8:40 a.m.	**Workshop Overview** Lynn Goldman, M.D., M.P.H. Roundtable Vice-Chair Dean, Milken Institute School of Public Health George Washington University
8:50 a.m.	**Session 1: Brief History and Current Legal and Executive Branch Framework for Data Sharing**
Objectives:	Federal rule making routinely involves the use of studies in which there are questions raised about how they were prepared and the reliability of the data and the conclusions drawn. Through the Administrative Procedure Act and the process of judicial review, there is now a reasonably well-developed understanding of how agencies handle those problems and what information must be made available to the public, to the Office of Information and Regulatory Affairs of the Office of Management and Budget, and to the courts (if judicial review is sought as it is for most significant rules).

The goal of this session is to inform workshop attendees on the basic ground rules of agency decision making in this area and to set the stage for the panels that follow.

Moderator: **Alan Morrison, LL.B.,** Lerner Family Associate Dean for Public Interest and Public Service Law, George Washington University

8:50 a.m. **Paul R. Verkuil, J.S.D.**
Chair, Administrative Conference of the United States

9:10 a.m. **George Gray, Ph.D.**
Professor
Department of Environmental and Occupational Health
Director of the Center for Risk Science and Public Health
Milken Institute School of Public Health
George Washington University

9:30 a.m. **Discussion** (20 minutes)

9:50 a.m.	**Session 2: Benefits and Importance of Data Sharing**

Objectives: Most researchers and policy makers in environmental health agree that, as in other fields, research data should be shared as freely as necessary, especially when research data are underpinning regulations. Most also recognize that there need to be some limits on when and how data should be shared.

The goal of this session is to clarify when, why, and how data sharing is beneficial for the researcher, the regulatory agency, the participants, and the public. The session should distinguish between sharing data for reanalysis and sharing data for facilitation of replication.

Moderator:	**Gwen Collman, Ph.D.,** Director, Division of Extramural Research and Training, National Institute of Environmental Health Sciences, National Institutes of Health
9:50 a.m.	Panelist Presentations (30 minutes)
	Bernard Lo, M.D. President and CEO The Greenwall Foundation
	Francesca Dominici, Ph.D. Professor of Biostatistics and Senior Associate Dean for Research Harvard University School of Public Health
	Julia Brody, Ph.D. Executive Director Silent Spring Institute
	Louis Anthony (Tony) Cox, Jr., Ph.D. Chief Sciences Officer, NextHealth Technologies
10:20 a.m.	**Panel Discussion** (60 minutes)

11:20 a.m.	**Session 3: Challenges Associated with Data Sharing**
Objectives:	While data sharing is quickly becoming the norm, especially for federally funded research, challenges exist for the investigator and her or his institution, the person or entity requesting the data, and the research participants, whose identity can be compromised. Concerns of the investigator are ensuring the right to publication and to quality control over reanalyses.

The goal of the session is to articulate the potential administrative, ethical, financial, and public health drawbacks to data sharing.

Moderator: **Glenn Paulson, Ph.D.,** Science Adviser, Office of the Administrator, U.S. Environmental Protection Agency

11:20 a.m. Panelist Presentations (30 minutes)

Daniel Barth-Jones, Ph.D., M.P.H.
Assistant Professor of Clinical Epidemiology
Mailman School of Public Health
Columbia University

John Howard, M.D., M.P.H., J.D., LL.M.
Director, National Institute for Occupational Safety and Health

Greg Bond, Ph.D.
Dow Masters Fellowship Program
University of Michigan

Kevin Casey, J.D.
Associate Vice President for Public Affairs and Communications
Harvard University

11:50 a.m. **Panel Discussion** (60 minutes)

12:50 p.m. **Lunch Break** (60 minutes)

1:50 p.m.	**Session 4: Ways Forward—Practices, Technologies, and Tools for Data Sharing**

Objectives: As data sharing becomes more common and an accepted practice, tools and best practices need to be identified to ensure the benefits of data sharing and avoid the challenges, as discussed earlier in the program.

	The goal of this session is to identify principles, practices, and programs that could be used more widely for maximizing data sharing.
Moderator:	**Lynn Goldman, M.D., M.P.H.,** Roundtable Vice-Chair
1:50 p.m.	Panelist Presentations (30 minutes)
	Edward Sondik, Ph.D. Former Director National Center for Health Statistics Centers for Disease Control and Prevention
	Daniel Greenbaum, M.S. President Health Effects Institute
	Linda S. Birnbaum, Ph.D., DABT, ATS Director National Institute of Environmental Health Sciences National Institutes of Health
	George Daston, Ph.D. Victor Mills Society Research Fellow Procter & Gamble
2:20 p.m.	**Panel Discussion** (60 minutes)
3:20 p.m.	**Break** (20 minutes)
3:40 p.m.	**Session 5: Reflections on the Workshop and Concluding Remarks**
Objective:	The objective of this session is to have reflections on the day, including next steps, from experts with different perspectives.
Moderator:	**Frank Loy, LL.B.,** Roundtable Chair

3:40 p.m.	**Linda S. Birnbaum, Ph.D., DABT, ATS** Director National Institute of Environmental Health Sciences National Institutes of Health
3:50 p.m.	**Jerry Blancato, Ph.D.** Director Office of Science and Information Management Office of Research and Development U.S. Environmental Protection Agency
4:00 p.m.	**Joseph Rodricks, Ph.D., DABT** Principal, ENVIRON
4:10 p.m.	**Ellen Silbergeld, Ph.D.** Editor in Chief *Environmental Research* Professor Johns Hopkins Bloomberg School of Public Health
4:20 p.m.	**Latanya Sweeney, Ph.D.** Professor of Government and Technology in Residence Harvard University
4:30 p.m.	**Discussion** (30 minutes)
5:00 p.m.	**Closing Remarks and Adjourn**

B

Speaker Biographical Sketches

Daniel C. Barth-Jones, Ph.D., M.P.H., is an infectious disease epidemiologist who specializes in computer simulation of the transmission and public health control of HIV and other infectious disease epidemics. His primary research interests include the epidemiology of HIV and sexually transmitted diseases, theoretical population vaccinology, Phase III HIV vaccine trial design, and health economic evaluations of public health policies for vaccination and preventative intervention programs. Dr. Barth-Jones is also a nationally recognized expert in the area of statistical disclosure analysis and control, where his work focuses on the development of statistical and geospatial disclosure control methodologies to help ensure the confidentiality and privacy of health care data in compliance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. He has given scientific presentations and conducted educational training on HIPAA privacy regulations to numerous health care information organizations, health care delivery organizations, and state and federal agencies and organizations and within academia. He has an M.P.H. and a Ph.D. from the University of Michigan School of Public Health.

Linda S. Birnbaum, Ph.D., DABT, ATS, became director of the National Institute of Environmental Health Sciences (NIEHS), one of the institutes of the National Institutes of Health, and the National Toxicology Program (NTP) in 2009. She is the first toxicologist and the first woman to lead NIEHS and NTP. In these roles, Dr. Birnbaum oversees federal funding for biomedical research to discover how the environment influences human health and disease. She has spent most of her career as a federal scientist. Her research and many of her publications focus on the pharmacokinetic behavior of environmental chemicals; the mechanisms of actions of toxicants, including endocrine disruption; and the linking of

real-world exposures to health effects. Dr. Birnbaum has received numerous awards and recognitions, including being elected to the National Academy of Medicine in October 2010. She also finds time to mentor the next generation of environmental health scientists as adjunct professor in the Gillings School of Global Public Health, the Curriculum in Toxicology, and the Department of Environmental Sciences and Engineering at the University of North Carolina at Chapel Hill, as well as in the Integrated Toxicology Program at Duke University.

Jerry Blancato, Ph.D., has been with the Office of Research and Development (ORD) of the U.S. Environmental Protection Agency (EPA) since joining EPA in 1985. He came to EPA after teaching for 10 years in Delaware. Dr. Blancato initially worked in what is now the National Center for Environmental Assessment in Washington, DC, researching and applying new methods for extrapolating doses between species. In 1989 he joined what is now the National Environmental Research Laboratory (NERL) in Las Vegas, Nevada, specializing in developing physiological pharmacokinetic models for use in exposure and risk assessment. After several years as a researcher, he became branch chief in 1997, acting division director in 2001, and acting associate director of gealth for NERL in 2004. He accepted a permanent position as the deputy director of the National Center for Computational Toxicology in 2005. In March 2009, he became the acting director of the Office of Administrative and Research Support. In July 2012, he became the acting director of the Office of Science Information Management (OSIM), and in September 2013 he was selected as OSIM's permanent director. He has served on numerous agency, interagency, and ORD committees, work groups, and task forces. In 2006 and 2007 he was cochair of the Information Technology Improvement Project and worked with numerous managers and staff members during the reorganization that eventually led to the formation of OSIM. Dr. Blancato has a bachelor's of science in chemistry, a master's of science in pathology, and a doctorate in biomedical engineering specializing in pharmacokinetic modeling.

Greg Bond, Ph.D., M.P.H., adjunct professor in the Department of Environmental Health in the University of Michigan School of Public Health, is a member of the team that is developing and delivering programming for the University of Michigan's Dow Sustainability Fellows Program, which will develop and support 300 sustainability scholars over the first 6 years of the program. Dr. Bond is on loan to the University of

Michigan from the Dow Chemical Company, where he most recently served as corporate director of product responsibility. Dr. Bond has published more than 60 peer-reviewed journal articles on epidemiology research and product stewardship. In 1988, he was elected a fellow in the American College of Epidemiology. He has served as an adviser to various government agencies, including the U.S. Environmental Protection Agency, the National Institute for Occupational Safety and Health, the National Cancer Institute, the Agency for Toxic Substances and Disease Registries, the Centers for Disease Control and Prevention, and others. Dr. Bond currently cochairs the International Council of Chemical Association's Chemicals Policy and Health Leadership Group, which is working to improve the product safety performance of the global chemical industry and its reputation and strengthen science and risk-based chemicals management legislation and regulation throughout the world. Dr. Bond has a Ph.D. in epidemiology and an M.P.H. from the University of Michigan.

Julia G. Brody, Ph.D., is a leader in research on breast cancer and the environment and in community-based research and public engagement in science. Dr. Brody's current research focuses on methods for reporting to people on their own exposures to hormone disrupters and other emerging contaminants when the health effects are uncertain. She recently led a project connecting breast cancer advocacy and environmental justice in a study of household exposures to endocrine disrupters and air pollutants through a collaboration of Silent Spring Institute, Communities for a Better Environment (a California-based environmental justice organization), and researchers at Brown University and the University of California, Berkeley. Since 1996, Dr. Brody has been the principal investigator of the Cape Cod Breast Cancer and Environment Study, a case-control study of 2,100 women that includes testing for 89 endocrine disrupters in homes and historical exposure mapping. Dr. Brody's research is supported by the National Science Foundation, the National Institutes of Health, the New York Community Trust, and the Avon Foundation, among others. The U.S. Environmental Protection Agency recognized her research with an Environmental Merit Award in 2000, and she has been honored by the Heroes Tribute of the Breast Cancer Fund. She serves on the National Advisory Environmental Health Sciences Council, to which she was appointed by the Secretary of Health and Human Services, and she is as an adviser to the California Breast Cancer Research Program and breast cancer activist organizations. Dr. Brody is an adjunct assistant professor at the

Brown University School of Medicine. She earned a Ph.D. at the University of Texas at Austin and an A.B. at Harvard University.

Kevin Casey, J.D., is associate vice president for public affairs and communications, and is responsible for the day-to-day administrative operations of the Department of Public Affairs and Communications and all of Harvard's government relations activities in Washington, DC. He also oversees representations on Harvard's behalf to the state of Massachusetts in regulatory and legislative matters. He specializes in legislative and regulatory issues relating to basic research, intellectual property, technology transfer, and many miscellaneous matters, including immigration policy. He has been with Harvard since April 1989. Prior to joining Harvard, Mr. Casey served as the Boston office chief of staff to Massachusetts Congressman Edward J. Markey. Earlier Mr. Casey was the staff director of the Massachusetts State Legislature's Joint Committee on Commerce and Labor. Mr. Casey is a graduate of Merrimack College and the New England School of Law and was admitted to the Massachusetts Bar in 1989.

Gwen Collman, Ph.D., leads approximately 60 professional staff in areas of scientific program administration, peer review, and the management and administration of about 1,500 active grants each year. She directs scientific activities across the field of environmental health sciences, including basic sciences (i.e., DNA repair, epigenetics, environmental genomics), organ-specific toxicology (i.e., reproductive toxicology, neurotoxicology, respiratory toxicology), public health–related programs (i.e., environmental epidemiology, environmental public health), and training and career development. She also oversees the implementation of the Superfund Research Program and the Worker Education and Training Program. Prior to her current role, Dr. Collman served in program development and management, beginning in 1992 as a member and then as chief of the Susceptibility and Population Health Branch. During this time, she directed research on the role of genetic and environmental factors on the development of human disease, from animal models of genetic susceptibility to population studies focusing on etiology and intervention. She was responsible for building the National Institute of Environmental Health Sciences (NIEHS) grant portfolio in environmental and molecular epidemiology and developed several complex multidisciplinary research programs. These include the NIEHS Breast Cancer and the Environment Research Centers Program, the NIEHS-U.S. Environmental Protection Agency Centers for Children's Environmental Health and Disease Pre-

vention, and the Genes, Environment, and Health Initiative. Also, under her guidance, a team created a vision for the Partnerships for Environmental Public Health programs for the next decade.

Louis Anthony (Tony) Cox, Jr., Ph.D., is president of Cox Associates, a Denver, Colorado-based applied research company specializing in quantitative risk analysis, causal modeling, advanced analytics, and operations research. Since 1986, Cox Associates' mathematicians and scientists have applied computer simulation and biomathematical models, statistical and epidemiological risk analyses, causal data mining techniques, and operations research and artificial intelligence models to measurably improve health, business, and engineering risk analysis and decision making for public- and private-sector clients. Since 1996, its sister company, NetAdvantage, has provided operations research services and software for telecommunications companies. He is also the chief sciences officer for NextHealth Technologies, a health care analytics software a service platform and services provider that leverages big data and prescriptive analytics to recommend, manage, and optimize consumer engagement. In 2006, Cox Associates was inducted into the Edelman Academy of the Institute for Operations Research and Management Science, which recognizes outstanding real-world achievements in the practice of operations research and the management sciences. In 2012, Dr. Cox was inducted into the National Academy of Engineering. He is a member of the National Academies of Sciences, Engineering, and Medicine's Board on Mathematical Sciences and Their Applications and a member of the Academies' Standing Committee on the Use of Public Health Data in Food Safety and Inspection Service food safety programs. Dr. Cox holds a Ph.D. in risk analysis and an S.M. in operations research, both from the Massachusetts Institute of Technology, and an A.B. from Harvard University and is a graduate of the Stanford Executive Program.

George Daston, Ph.D., is a Victor Mills Society Research Fellow at the Procter & Gamble Company. His current research efforts are in the areas of toxicogenomics and mechanistic toxicology, particularly in addressing how findings in these fields can improve risk assessment for chemicals and the development of nonanimal alternatives to testing. He has published more than 100 articles and book chapters and edited 5 books in toxicology and risk assessment. Dr. Daston has served as president of the Teratology Society, as councilor of the Society of Toxicology, on the Board of Scientific Counselors of the U.S. Environmental Protection Agency, the

National Toxicology Program Board of Scientific Counselors, the National Academies of Sciences, Engineering, and Medicine's Board of Environmental Studies and Toxicology, and the National Children's Study Advisory Committee. He is editor in chief of *Birth Defects Research: Developmental and Reproductive Toxicology*. Dr. Daston manages the AltTox website, which is devoted to the exchange of scientific information leading to the development of in vitro replacements for toxicity assessments. Dr. Daston has been awarded the Josef Warkany Lectureship by the Teratology Society and the George H. Scott Award by the Toxicology Forum and was elected a fellow of the American Association for the Advancement of Science. Dr. Daston is an adjunct professor of pediatrics at the University of Cincinnati.

Francesca Dominici, Ph.D., is senior associate dean for research and professor of biostatistics at the Harvard University School of Public Health. Dr. Dominici has authored more than 120 peer-reviewed publications. Her research has focused on the development of statistical methods for the analysis of large observational data with the ultimate goal of addressing important questions in environmental health science, the health-related impacts of climate change, and comparative effectiveness research. She is an expert in Bayesian methods, longitudinal data analysis, confounding adjustment, causal inference, and Bayesian hierarchical models. She has extensive experience with the development of statistical methods and their applications to environmental epidemiology, implementation science and health policy, outcomes research and patient safety, and comparative effectiveness research. Dr. Dominici received a Ph.D. in statistics from the University of Padua, Italy. During her Ph.D. studies, she spent 2 years as a visiting student at Duke University. Afterward, she attended the Bloomberg School of Public Health at Johns Hopkins University as a postdoctoral fellow.

Lynn R. Goldman, M.D., M.S., M.P.H., is a world-renowned epidemiologist, pediatrician, educator, and former regulator of the U.S. Environmental Protection Agency (EPA). She was named dean of the George Washington University School of Public Health and Health Services, as it was known then in 2010. In 2014, she assumed the Michael and Lori Milken Dean of Public Health at the newly renamed Milken Institute School of Public Health at George Washington University. Her areas of focus are public health practice, children's environmental health, disaster preparedness, and chemical and pesticide regulatory policy. As

assistant administrator for toxic substances at EPA, she directed the Office of Prevention, Pesticides, and Toxic Substances from 1993 through 1998. Prior to joining EPA, Dr. Goldman served as chief of the Division of Environmental and Occupational Disease Control of the California Department of Health Services. Dr. Goldman has served on numerous boards and expert committees, including the Committee on Environmental Health of the American Academy of Pediatrics and the Centers for Disease Control and Prevention Lead Poisoning Prevention Advisory Committee. Dr. Goldman is a member of the National Academy of Medicine, vice chair of the National Academies of Sciences, Engineering, and Medicine Roundtable on Environmental Health Sciences, Research, and Medicine, and a member of the Academies' Standing Committee on Risk Analysis Issues and Reviews.

George Gray, Ph.D., has long been committed to the effective use of science to inform public health choices in both academic and policy-making settings, and emphasizes the importance of effectively communicating those choices to citizens, journalists, and lawmakers. Prior to joining George Washington University in 2010, Dr. Gray served as assistant administrator for the U.S. Environmental Protection Agency's (EPA's) Office of Research and Development and as the Agency's science adviser, promoting scientific excellence in EPA research, advocating for the continuing evolution of the Agency's approach to analysis, and encouraging programs that provide academic research to support EPA's mission. His areas of focus included nanotechnology, ecosystem research, the influence of toxicology advances on testing and risk assessment, and sustainability. From 2001 to 2005, Dr. Gray was executive director of the Harvard University Center for Risk Analysis and a member of the faculty at the Harvard University School of Public Health. In addition to teaching, he applied the tools of risk analysis to public health problems ranging from mad cow disease to pesticides in food and the risks and benefits of fish consumption. He has an M.S. in toxicology and a Ph.D. from the University of Rochester School of Medicine and Dentistry.

Daniel Greenbaum, M.S., joined the Health Effects Institute (HEI) as president and chief executive officer in 1994. In that role, Greenbaum leads HEI's efforts, supported jointly by the U.S. Environmental Protection Agency (EPA) and industry, with additional funding from the U.S. Department of Energy, the Federal Highway Administration, the U.S.

Agency for International Development, the Asian Development Bank, and foundations, to provide public and private decision makers with high-quality, impartial, relevant, and credible science about the health effects of air pollution. Mr. Greenbaum has focused HEI's efforts on providing timely and critical research and reanalysis on particulate matter, air toxics, diesel exhaust, and alternative technologies and fuels. Mr. Greenbaum currently serves on the National Academies of Sciences, Engineering, and Medicine Committee on Health, Environmental, and Other External Costs and Benefits of Energy Production and Consumption. He has been a member of the Academies Board on Environmental Studies and Toxicology and vice-chair of its Committee for Air Quality Management in the United States. Mr. Greenbaum also chaired the EPA Blue Ribbon Panel on Oxygenates in Gasoline, which issued the report Achieving Clean Air and Clean Water, and EPA's Clean Diesel Independent Review Panel, which reviewed technology progress in implementing the 2007 Highway Diesel Rule. Before coming to HEI, he was commissioner of environmental protection in Massachusetts.

John Howard, M.D., M.P.H., J.D., LL.M., serves as the director of the National Institute for Occupational Safety and Health (NIOSH) in the U.S. Department of Health and Human Services in Washington, DC. He served in this capacity from July 2002 to July 2008 and was reappointed in September 2009. Prior to his appointment as director of NIOSH, Dr. Howard served as chief of the Division of Occupational Safety and Health in the California Department of Industrial Relations from 1991 through 2002. Dr. Howard is board certified in internal medicine and occupational medicine. He is admitted to the practice of medicine and law in the state of California and in the District of Columbia, and he is a member of the U.S. Supreme Court bar. He has written numerous articles on occupational health law and policy. Dr. Howard received a doctor of medicine degree from Loyola University of Chicago, a master's of public health degree from the Harvard School of Public Health, a doctor of law degree from the University of California, Los Angeles, and a master of law in administrative law from the George Washington University.

Bernard Lo, M.D., was professor of medicine and director of the program in medical ethics at the University of California, San Francisco, before becoming president of The Greenwall Foundation. Currently he cochairs the Standards Working Group of the California Institute of Regenerative Medicine, which recommends regulations for stem cell

research funded by the state of California. At the National Institutes of Health (NIH), he serves on Data and Safety Monitoring Committees for HIV vaccine trials and the Long-Term Oxygen Treatment Trial. Dr. Lo serves on the board of directors of the Association for the Accreditation of Human Research Protection Programs and on the Medical Advisory Panel of Blue Cross/Blue Shield. He was formerly a member of the National Bioethics Advisory Commission under President Bill Clinton, the NIH Recombinant DNA Advisory Committee and the Ethics Subcommittee, and the Advisory Committee to the Director of the Centers for Disease Control and Prevention. He has written articles on ethical issues in the procurement of embryos for research, oversight of stem cell lines derived in other institutions, informed consent for future research, and prohibiting the use of induced pluripotent stem cells for reproductive cloning. Dr. Lo is the author of *Resolving Ethical Dilemmas: A Guide for Clinicians* (5th ed., Lippincott Williams & Wilkins, 2013) and of *Ethical Issues in Clinical Research* (Lippincott Williams & Wilkins, 2010). A member of the National Academy of Medicine, Dr. Lo served on the National Academy of Medicine Council and chaired the Board on Health Sciences Policy. He chaired National Academies of Sciences, Engineering, and Medicine committees on conflicts of interest in medicine and on confidentiality in health services research and has been a member of several other Academies committees. He is currently a member of the Board on Life Sciences of the Academies.

Frank Loy, LL.B., has served in the U.S. Department of State in four administrations. He served as Undersecretary of State for Global Affairs in the second administration of President Bill Clinton. His portfolio included developing U.S. international policy and conducting negotiations in the fields of the environment and climate change, human rights, the promotion of democracy, refugees and humanitarian affairs, and counternarcotics. Under President Jimmy Carter he was director of the Bureau of Refugee Programs with the personal rank of ambassador, and in the Lyndon Johnson Administration he served as Deputy Assistant Secretary for Economic Affairs. In 2011 President Barack Obama named him the U.S. Alternate Representative to the United Nations General Assembly. At present he serves on the boards of numerous nonprofit organizations. In the field of the environment these include Resources for the Future (former chair), the Environmental Defense Fund (former chair), The Nature Conservancy, the Center for Climate and Energy Solutions, and ecoAmerica (chair). He also chairs the boards of Population Services International and the Arthur Burns

Fellowship Program and serves on the boards of the American Institute for Contemporary German Studies and The Washington Ballet.

Alan B. Morrison, LL.B., is the Lerner Family Associate Dean for Public Interest & Public Service at the George Washington University Law School. Morrison graduated from Yale University and the Harvard University Law School. He also worked as an assistant U.S. attorney in the Southern District of New York. For most of his career he was with the Public Citizen Litigation, which he cofounded with Ralph Nader and directed for many years. He has taught administrative law and other subjects at Harvard, New York University, Stanford University, the University of Hawaii, and American University law schools. He is a senior fellow of the Administrative Conference of the United State and a member of the National Academies of Sciences, Engineering, and Medicine's Committee on Science, Technology, and Law.

Glenn Paulson, Ph.D., is co-chair of the Committee on Environment, Natural Resources, and Sustainability of the National Science and Technology Council, an operating unit of the White House Office of Science and Technology Policy. Dr. Paulson has more than 40 years' experience in environmental science, technology, and policy issues in the public, private, and academic sectors. He served three terms on the National Academies of Sciences, Engineering, and Medicine's Nuclear and Radiation Studies Board and has been a member of many federal government advisory committees, including a charter member of the U.S. Secretary of Energy's Advisory Board and the first chair of the U.S. Department of Energy's Environmental Management Advisory Board. In his current position as science adviser to the administrator of the U.S. Environmental Protection Agency (EPA), Dr. Paulson provides expertise on a wide variety of scientific issues relevant to Agency decisions. He also chairs the cross-agency Science and Technology Policy Council, which is made up of high-level managers from across EPA. The Council deals with issues such as the Laboratory Enterprise Study, responses to recommendations from Academies reports, technology innovation, sustainability, and improving risk assessment methods and approaches, as well as environmental measurement, modeling, and monitoring. Dr. Paulson also serves as EPA's acting scientific integrity official and leads the implementation of EPA's policy on scientific integrity. He is also responsible for overseeing EPA's compliance with the requirements for human subjects research.

Joseph V. Rodricks, Ph.D., DABT, is a founding principal of ENVIRON and an internationally recognized expert in toxicology and risk analysis. He has consulted for hundreds of manufacturers, government agencies, and the World Health Organization in the evaluation of health risks associated with human exposure to chemical substances of all types. Dr. Rodricks came to consulting after a 15-year career as a scientist at the U.S. Food and Drug Administration (FDA). In his last 4 years at FDA, he served as associate commissioner for health affairs. His experience extends from pharmaceuticals, medical devices, consumer products, and foods to occupational chemicals and environmental contaminants. He has served on the National Academies of Sciences, Engineering, and Medicine's Board on Environmental Studies and Toxicology and on 30 boards and committees of the Academies, including the committees that produced the seminal works *Risk Assessment in the Federal Government: Managing the Process* (1983) and *Science and Decisions: Advancing Risk Assessment* (2009). He has published nearly 150 scientific publications and has received honorary awards from 3 professional societies for his contributions to toxicology and risk analysis. He is author of the widely used text *Calculated Risks*, now in its second edition, published by Cambridge University Press.

Ellen Silbergeld, Ph.D., trained at Johns Hopkins University in geography and environmental engineering and is a postdoctoral fellow in environmental health sciences. She had a staff fellowship at the National Institutes of Health (NIH) followed by senior scientist position at Environmental Defense Fund and a professorship at the University of Maryland Medical School. She has served as scientific adviser to the National Toxicology Program of NIH, the Centers for Disease Control and Prevention, the U.S. Environmental Protection Agency, the U.S. Department of Energy, the Occupational Safety and Health Administration, the states of Maryland and New York, the World Bank, the International Labour Organization, the United Nations Environment Programme, and the Pan American Health Organization. Her research and professional activities bridge science and public policy and a focus on the incorporation of mechanistic toxicology into environmental and occupational health policy. Areas of current focus include the cardiovascular risks of arsenic, lead, and cadmium; the immunotoxicity of mercury compounds; and the health and environmental impacts of industrial food animal production. These projects include epidemiological studies and mechanistic research on gene–environment interactions and

the movement of pathogens in the environment. Some of this research is conducted internationally (mercury studies in the Amazon; lead/ cadmium/arsenic studies in Mexico; mining and development studies in Mongolia; zoonotic disease studies in Thailand and the Netherlands). She also directs a Fogarty Training Program in noncommunicable diseases, which is a collaboration between Johns Hopkins University and the School of Public Health of Mongolia.

Edward J. Sondik, Ph.D., is director of the National Center for Health Statistics, directing personnel located in two locations: in Hyattsville, Maryland, outside Washington, DC, and in Research Triangle Park, North Carolina. He served as director since joining the Centers for Disease Control and Prevention in 1997. Dr. Sondik has also served as the acting director of the National Center for Public Health Informatics and as acting director to co-lead the Coordinating Center for Health Information and Service until a candidate was named. Dr. Sondik's background is in mathematics and statistics and the discipline of operations research and has academic training in electrical engineering on the side of control systems, computers, and operations research. Dr. Sondik received a Ph.D. at Stanford University. It was there that he was given the opportunity to work with the Stanford Medical School to redesign the Stanford Hospital. He also taught at the university.

Latanya Sweeney, Ph.D., is professor of government and technology in residence at Harvard University and director and founder of the Data Privacy Lab, now at Harvard University. Prior to this she was a distinguished career professor of computer science, technology, and policy in the School of Computer Science at Carnegie Mellon University. Dr. Sweeney's work involves creating technologies and related policies with guarantees of privacy protection while allowing society to collect and share person-specific information for many worthy purposes. Her work has received awards from numerous organizations, including the American Psychiatric Association, the American Medical Informatics Association, and the Blue Cross Blue Shield Association. Dr. Sweeney received a Ph.D. in computer science from the Massachusetts Institute of Technology in 2001, being the first black woman to do so. Her undergraduate degree in computer science was from Harvard University.

Paul R. Verkuil, J.S.D., the 10th chairman of the Administrative Conference of the United States (ACUS), was sworn in by Vice President

Joe Biden on April 6, 2010. The Conference was revived by Congress in 2009 after a 15-year hiatus. President Barack Obama named the 10-member Council on July 8, 2010, saying, "ACUS is a public–private partnership designed to make government work better." The 50 government and 40 public members, along with the council and chair, form the 101-member Conference. The Conference meets twice per year in June and December in plenary sessions to make consensus-driven recommendations to improve government processes and procedures. Mr. Verkuil is a well-known administrative law teacher and scholar who has co-authored a leading treatise, *Administrative Law and Process*, now in its fifth edition; several other books (most recently, *Outsourcing Sovereignty*, Cambridge University Press, 2007); and more than 65 articles on the general topic of public law and regulation. He is president emeritus of the College of William & Mary, former dean of the Tulane and Cardozo Law Schools, and a faculty member at the University of North Carolina Law School. He is a graduate of the College of William & Mary and the University of Virginia Law School and holds a J.S.D. from New York University Law School. He is a life member of the American Law Institute and a fellow of the American Bar Foundation.