Sharing Research Data to Improve Public Health in Africa: A Workshop Summary

DETAILS

102 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-37809-3 | DOI 10.17226/21801

AUTHORS

Mary Ellen O'Connell and Thomas J. Plewes, Rapporteurs; Committee on Population; Division of Behavioral and Social Sciences and Education; The National Academies of Sciences, Engineering, and Medicine

Sharing Research Data to Improve Public Health in Africa

A Workshop Summary

Mary Ellen O'Connell and Thomas J. Plewes, Rapporteurs

Committee on Population

Division of Behavioral and Social Sciences and Education

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

Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project. This activity was supported by the National Institute on Aging of the National Institutes of Health through Contract No. 10001706, Order No. HHSN26300056.

International Standard Book Number-13: 978-0-309-37809-3
International Standard Book Number-10: 0-309-37809-5

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

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. (2015). *Sharing Research Data to Improve Public Health in Africa: A Workshop Summary.* M.E. O'Connell and T.J. Plewes, Rapporteurs. Committee on Population, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

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**.

STEERING COMMITTEE FOR A WORKSHOP ON STRENGTHENING SCIENCE TO INFORM PUBLIC HEALTH POLICY: OVERCOMING BARRIERS TO SHARING RESEARCH DATA IN SUB-SAHARAN AFRICA

DAVID CARR (*Chair*), Wellcome Trust
MUHAMMAD ALI DHANSAY, South Africa's Nutritional Intervention Research Unit, Medical Research Council
ROSEANNE DIAB, Academy of Science of South Africa
STEVEN E. KERN, Bill & Melinda Gates Foundation
ROBERT TERRY, World Health Organization

THOMAS J. PLEWES, *Study Director*
MARY ELLEN O'CONNELL, *Rapporteur*
MARY GHITELMAN, *Program Assistant*

COMMITTEE ON POPULATION
2015

KATHLEEN MULLAN HARRIS (*Chair*), Department of Sociology, University of North Carolina at Chapel Hill
JERE R. BEHRMAN, Department of Economics, University of Pennsylvania
VICKI A. FREEDMAN, Institute for Social Research, University of Michigan
MARK D. HAYWARD, Population Research Center, University of Texas at Austin
HILLARD S. KAPLAN, Department of Anthropology, University of New Mexico
SARA S. McLANAHAN, Center for Research on Child Wellbeing, Princeton University
EMILIO A. PARRADO, University of Pennsylvania
DAVID R. WEIR, Survey Research Center, Institute for Social Research, University of Michigan
JOHN R. WILMOTH, Population Division/DESA, United Nations

THOMAS J. PLEWES, *Director*
TINA M. LATIMER, *Program Coordinator*

Preface and Acknowledgments

This report summarizes a workshop convened in Stellenbosch, South Africa, on March 29–30, 2015, which focused on the benefits of and barriers to sharing research data in order to improve public health. The workshop was sponsored by the Wellcome Trust and the Bill & Melinda Gates Foundation. This workshop summary report was sponsored by the National Institute on Aging (NIA) of the National Institutes of Health and is a product of the Committee on Population of the National Academies of Sciences, Engineering, and Medicine (the Academies).

The purposes of the workshop were to raise the profile of issues around the sharing of public health data in Africa, enable the Wellcome Trust and its international partners to highlight findings of previous sponsored research on this topic, identify issues that mitigate against public health data sharing and pathways through research and policy venues to foster increased sharing, and, in general, serve as a way to bring more African voices and perspectives into the dialogue. It was conducted in cooperation with several sponsoring organizations and representatives of national science academies in Africa, as well as experts in using and generating public health data to discuss the benefits of and barriers to sharing research data within the African context.

The workshop was organized by a committee of experts representing several of the sponsoring organizations. The committee was chaired by David Carr, policy adviser, Wellcome Trust, and included Muhammad Ali Dhansay, director, South Africa's Nutritional Intervention Research Unit, Medical Research Council; Roseanne Diab, executive director, Academy

of Science of South Africa; Steven Kern, deputy director of quantitative sciences, Bill & Melinda Gates Foundation; and Robert Terry, senior strategic and project manager, World Health Organization. Georgeanne E. Patmios of NIA and Kobus Herbst of INDEPTH were also members of the committee that organized the conference. The committee provided guidance in developing the workshop agenda, secured expert presentations, and facilitated the conduct of the workshops. The meeting was hosted by the Academy of Science of South Africa and the South African Medical Research Council, whose representatives also served as members of the organizing committee. Although the steering committee members played a central role in planning and conducting the workshop, they did not actively participate in writing this summary. The committee benefited from the active participation of the late Richard Suzman, NIA, in the planning phase and from the support of Georgeanne E. Patmios of NIA in the workshop itself.

The presentations during the workshops provided the basis for lively and informative discussions. As summarized in this report, each of the five sessions was introduced in a keynote presentation by an acknowledged expert in the subject matter, followed by individual or panel presentations. The contributions of the session chairs, keynote speakers, and presenters—identified in the agenda that appears as Appendix A to this report—are gratefully acknowledged.

The steering committee acknowledges the work of the staff of the Academies in organizing the workshops and this summary. The Committee on Population provided overall direction and guidance for the project, and Mary Ellen O'Connell and Thomas Plewes served as rapporteurs for this report. Mary Ghitelman provided exceptional assistance with administrative and logistical arrangements, and in the production of this summary report. Paula Whitacre ably edited the report, and Kirsten Sampson Snyder and Eugenia Grohman orchestrated the review and editing processes.

This workshop summary was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the Academies. The purpose of this independent review is to provide candid and critical comments that assist the institution in making its report as sound as possible, and to ensure that the report 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 deliberative process.

The panel thanks the following individuals for their review of this report: David Carr, policy adviser, Wellcome Trust; Pierre Ongolo-Zogo, coordinator, Specialized Internship Program University of Yaoundé and

director, Centre for Development of Best Practices in Health, Yaoundé, Cameroon; and Osman Sankoh, executive director, INDEPTH network, Accra, Ghana.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by William Eddy, Department of Statistics, Carnegie Mellon University. Appointed by the Academies, he was responsible for making certain that the independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of the report rests entirely with the author and the National Academies of Sciences, Engineering, and Medicine.

Thomas J. Plewes
Director
Committee on Population

Contents

1

Introduction

Sharing research data on public health issues can promote expanded scientific inquiry and has the potential to advance improvements in public health. Although sharing data is the norm in some research fields, such as the social sciences, sharing of data in public health is not as firmly established. On March 29–31, 2015, representatives of the Wellcome Trust, Bill & Melinda Gates Foundation, National Institute on Aging, Special Programme for Research and Training in Tropical Diseases of the World Health Organization, INDEPTH network, South Africa Medical Research Council, and Academy of Sciences of South Africa organized a workshop in Stellenbosch, South Africa, to explore issues related to sharing research data to improve public health in an African context. Hosted by the South African Medical Research Council and the Academy of Sciences of South Africa, the workshop brought together public health researchers and epidemiologists primarily from the African continent, along with selected international experts, to talk about the benefits and challenges of sharing data to improve public health, and to discuss potential actions to guide future work related to public health research data sharing.

In the course of five major sessions, each characterized by a keynote presentation and ample time for panel and floor discussions, the workshop participants discussed many issues that are detailed further in this summary:

- *There is growing international support for data sharing.* The public health benefits of research conducted with shared data have been demonstrated in multiple settings. Funders of public health research, including members of the Public Health Research Data Forum, have begun requiring data sharing. Similarly, the United States has begun implementing a national requirement that data generated by federally funded research be shared. A long history of data sharing exists in the social sciences. Some journals are also now requiring that data be shared as a condition of publication; other journals (e.g., Ubiquity Press) are developing approaches to allow data to be published. In the course of the meeting, several examples of collaborations that fostered an environment for data sharing were offered.
- *Data-sharing issues are not unique to Africa, but context matters.* The issues and concerns tied to data sharing raised at the workshop (e.g., confidentiality, data quality, community relevance, cost, and ownership) are relevant to data sharing in a general public health context. However, the historical context of exploitation in Africa, the power imbalance resulting from the tendency for data collectors to be in Africa and analysts to be in the Global North, and the lack of infrastructure, combined with resource inequities that exacerbate this lack, raise a unique situation for resolving these issues in the African context. Familiarity with data sharing and data-sharing issues also appears to be limited in low- and middle-income countries, including African countries. This may contribute to nervousness among African researchers with sharing data outside of established collaborations. It also highlights the need for equity and fairness in research contracts.
- *Data from Africa should benefit Africa.* Many participants at the workshop conveyed a sentiment that data generated from Africa should result in a benefit to Africa—not just in terms of public health generally, but for the data subjects, African researchers, and African institutions involved. Collaborations such as H3Africa, INDEPTH, and the ALPHA network provide useful illustrations of how data sharing can be used in health policy and models to address data-sharing issues.
- *Sharing has both risks and benefits.* Sharing of data presents both risks and benefits, or challenges and opportunities, to the individual providing the data, to the researcher, to the institution, and to the community. Developing a framework for data sharing requires an appropriate balance of the relevant risks and benefits.
- *Data sharing exists within a data cycle continuum.* To have data of sufficient quality and quantity to enable them to be shared and

productively used in secondary research requires attention to issues of data collection, standardization, curation, and management, including the associated costs. Attention to these issues requires consideration of the roles, expectations, and benefits of those involved at each stage of the research process, and for the roles, expectations, and benefits to be articulated at the start of a project.

- *Data sharing is a crucial element of the research continuum.* Sharing data at the end of a research project is an essential part of the process. Collecting good, standardized data that can be shared should be viewed as good scientific practice.

2

Context

David Carr, policy adviser at the Wellcome Trust and chair of the National Research Council (NRC) Steering Committee for a Workshop on Strengthening Science to Inform Public Policy, opened the workshop. He highlighted the interests of funders of public health research and of the Public Health Research Data Forum (PHRDF),[1] which he manages, in sharing research data to improve public health. The PHRDF members, most of whom now require data sharing as a condition of funding, believe that making research data available to researchers beyond those who originally collected the data will lead to faster progress in improving health, better value for the invested resources, and higher-quality science overall, he explained.

Despite general agreement that sharing of data has the potential to generate both research and policy benefits, putting that agreement into practice is not a simple matter, Carr observed. Standard considerations such as privacy and confidentiality are of concern in all data-sharing

[1]Members of the PHRDF include the Agency for Healthcare Research and Quality (U.S.), Bill & Melinda Gates Foundation, Canadian Institutes of Health Research, Centers for Disease Control and Prevention (U.S.), Deutsche Forschungsgemeinschaft, Doris Duke Charitable Foundation, Economic and Social Research Council (UK), Human Research Council of New Zealand, Health Resources and Services Administration (U.S.), Hewlett Foundation, Institut National de la Santé et de la Recherche Médicale (France), Medical Research Council (UK), National Health and Medical Research Council (Australia), National Institutes of Health (U.S.), Substance Abuse and Mental Health Services Administration (U.S.), U.S. Agency for International Development, Wellcome Trust, and World Bank.

arrangements, he noted, but particular challenges exist when data are collected by researchers in low- and middle-income countries and shared with researchers in better-resourced research centers who enjoy the benefit of analyzing the data. These issues served as the basis for much of the discussion at the workshop.

Carr highlighted the interests of the PHRDF members in having data shared in a manner that is responsive to three principles:

1. *Equitable.* Data sharing should recognize and balance the needs of different communities involved, including those who generate the data, the communities from which the data came, secondary users of the data, and funders of the data collection effort.
2. *Ethical.* Data sharing should protect the privacy of individuals and the dignity of affected communities, while also ensuring the maximum benefit to public health through use of shared data.
3. *Efficient*. Data sharing should improve the quality and value of research, aim to improve public health, build on existing best practice, and avoid unnecessary duplication and competition.

Carr highlighted several initiatives being undertaken by PHRDF[2] to advance the goals of increased data sharing and improved public health. He also noted several trends that point to the importance of data sharing. For example, the UK Royal Society produced a report in 2012, *Science as an Open Enterprise*,[3] which pointed to the benefits. Funders have been supporting this value through policies mandating data sharing. Journals have also been increasingly vocal: For example, PLOS recently established a policy that requires that data underlying published research be shared. The 2013 statement by the G8 Science Ministers similarly explicitly highlighted the importance of access to both research data and research publications.[4]

Carr pointed to parallel developments tied to privacy and confidentiality protections. For example, the European Data Protection Regulations call for more stringent protections that could impede data sharing, and South Africa has passed a privacy law tied to human subject research. At the same time, in his view, there is growing interest in sharing clinical trial data, concerns about research reproducibility, attention to issues of dupli-

[2]For a description of the initiatives, see http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030689.htm [August 2015].

[3]See https://royalsociety.org/policy/projects/science-public-enterprise/Report/ [August 2015].

[4]See https://www.gov.uk/government/news/g8-science-ministers-statement [August 2015].

BOX 2-1
Statement of Work

An ad hoc committee, established under the auspices of the National Research Council's Standing Committee on Population and working in coordination with the Wellcome Trust, will organize an international conference in Africa on the benefits of and barriers to sharing research data in order to improve public health. The conference will involve the participation of representatives of national science academies in Africa as well as experts in using and generating public health data and will feature presentations and discussions on the benefits of and barriers to sharing research data within the African context. The conference will be held in a host location in Africa. The conference will afford an opportunity to raise the profile of this issue within Africa, enable the Wellcome Trust and its international partners to highlight findings of previous sponsored research on this topic, identify issues that mitigate against public health data sharing and pathways through research and policy venues to foster increased sharing, and, in general, serve as a way to bring more African voices and perspectives into the dialogue. The immediate product of the conference will be a rapporteur-authored summary that can help to inform the future course of public health data sharing in Africa.

cation and waste, and corresponding interest in maximizing efficiency. He said these concerns form the background for a discussion of data sharing.

Carr closed his comments by laying out the primary goal for the workshop—to articulate opportunities and challenges in relation to increasing the availability of health research data in the African context. The workshop built on relationships between the U.S. National Academy of Sciences and African Academies of Science. While data sharing poses issues that are not unique to Africa, a discussion within the African context was viewed as potentially valuable. The formal statement of work for the workshop can be found in Box 2-1.

BENEFITS AND CHALLENGES OF DATA SHARING IN THE AFRICAN CONTEXT: THE INDEPTH NETWORK

Kobus Herbst, deputy director of the Africa Center for Health and Population Studies, presented on the multiyear experience of the INDEPTH network,[5] a network of approximately 50 research centers that run health and demographic surveillance systems (HDSS), mostly in Africa but also in India, Southeast Asia, and Oceania. Demographic surveillance involves

[5]For additional information, see http://www.indepth-network.org [August 2015].

collecting information on a census of individuals in a geographically defined area, and then tracking information about them over time. It includes individuals born to residents within the cohort area as well as those who immigrate to the area; individuals are excluded when they move out of the area or die. Information is collected on measures such as characteristics of the environment of households and information about the individuals such as socioeconomic status, vaccines, HIV, nutrition, and the like. Interventions, randomized trials, and other health system interventions are conducted on the cohort populations and the outcomes of the interventions are evaluated using the surveillance data, as well as information available on disease episodes and hospital admissions through linkages with local health information. HDSS networks typically are in places with no vital statistics and represent the only information available on health status and processes in their communities.

Herbst described his early thinking on data access as being oriented toward the goal to balance three perspectives: data subjects, data collectors or producers, and data users (see Figure 2-1). Data subjects are concerned about confidentiality, how the data will or may be used, and how they might benefit, or at least not be harmed. Data collectors or producers want to produce high-quality data, attract and retain staff, protect respondents, and sustain their projects. Data users aim to advance science, answer new questions, and inform policies.

Over time, he said, the network grappled with questions about where data sharing is a good thing. For example, he questioned the quality of certain data and the capacity to manage the sharing process if it extended beyond the immediate network of collaborating scientists. Questions were also raised about what data to share and the mechanics of sharing the data at the network level, keeping in mind that data must be sufficiently robust to enable sharing. Finally, he said, the network had a sense that they "should not just blindly share data"; rather, they should promote a specific research agenda and the concept of "fair trade."

In 2008, an article in *PLOS Medicine*[6] highlighted the debate, Herbst pointed out. One side argued that suboptimal access to data impedes international research and the potential substantial benefits of sharing, while the other side argued major technical obstacles should be addressed. The paper also pointed to the pioneering work being done by the networks and the fact that "the developing country scientist wants to move away from being primary producers of data for developed country scientists

[6]Chandramohan, D., Shibuya, K., Setel, P., Cairncross, S., Lopez, A., Murray, C., Zaba, B., Snow, R., and Binka, F. (2008). Should data from demographic surveillance systems be made more widely available to researchers? *PLOS Medicine*, 5(2), E57. See http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050057 [August 2015].

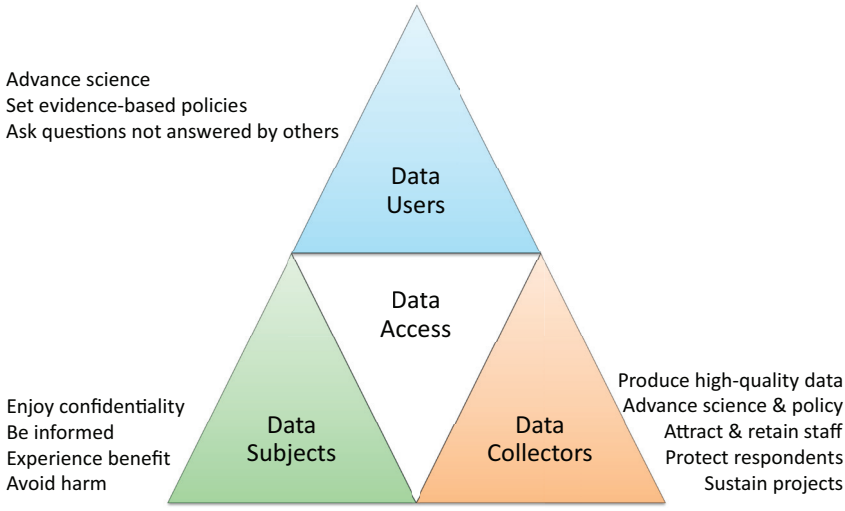

FIGURE 2-1 Three perspectives on data access.
SOURCE: Adapted from Herbst, K. (2002). Wider accessibility to longitudinal datasets: A framework for discussion. In National Research Council, *Leveraging Longitudinal Data in Developing Countries: Report of a Workshop* (p. 43). Workshop on Leveraging Longitudinal Data in Developing Countries Committee, Committee on Population. V. L. Durrant and J. Menken, Eds. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

to analyze. They do not wish to remain hewers of data and drawers of protocols. There's an urgent need to enable scientists in the South to play an equal role in the analysis of data they gather to support the national governments in the science policy interface and to develop science careers through appropriate citation in internationally peer-reviewed journals."

This introduction of the concept of "fair trade" in data sharing has continued. Osman Sankoh, executive director of INDEPTH, and Carel Ijsselmuiden, director of the Council on Health Research for Development South Africa, published an opinion, Sharing Research Data to Improve Public Health, A Perspective from the Global South in *The Lancet*,[7] which stated that fair trade in data "implies achieving a balance between the rights and responsibilities of those who generate data and those who analyze and publish results using those data. Such a balance lies in ensur-

[7]Sankoh, O., and Ijsselmuiden, C. (2011). Sharing research data to improve public health: A perspective from the Global South. *The Lancet, 378*(9789), 401–402, 430. See http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(11)61211-7/fulltext?rss=yes [August 2015].

ing that the means and capacity to share and actively participate in the analysis of those data are in the hands of those who generate the data and not only in those who want to" (p. 401).

In 2007, three INDEPTH sites published their data in a self-developed repository. In 2008, three additional African sites joined, and the repository was later expanded to include sites in Vietnam, Bangladesh, and Nairobi, Kenya. To responsibly enable sharing of these data, INDEPTH developed a data access and sharing policy agreed to and owned by all INDEPTH members. The policy was published in the *International Journal of Epidemiology* in April 2011, together with the first specification of a standard micro dataset that the network would share.[8]

In July 2013, they launched the INDEPTH Data Repository and an indicator site called INDEPTH Stats. Since then, about 30 sites have extracted quality-assured data; twelve of those datasets are in the data repository and the rest of them are being curated for placement in the repository. *PLOS ONE* now recognizes the INDEPTH Data Repository as an acceptable repository for its publications.

Challenges

Herbst shared some challenges that INDEPTH encountered during the process:

- *Data harmonization*. Because sites had different levels of information technology available, different databases, and different data structures, there was significant work done to harmonize data and make available to sites a consistent environment within which data could be extracted and documented.
- *Research data management skills*. Most sites struggle with limited research data management skills. It is hard to hire and retain staff with the needed skills.
- *Data quality*. Maintaining data quality over many years of longitudinal individual-level surveillance is a challenge, particularly when dealing with highly mobile populations where there are no unique individual identifiers.
- *Conceptual differences*. Involvement of different countries introduces potential differences in basic concepts, such as the definition of a household or the unit of representation of social groups. Publishing shared data on a household dynamic analysis requires

[8]Sankoh, O., and Byass, P. (2012). The INDEPTH network: Filling vital gaps in global epidemiology. *International Journal of Epidemiology, 41*(3), 579–588. See http://ije.oxfordjournals.org/content/41/3/579.full [August 2015].

a common definition. Similarly, it requires agreement on definitions of educational attainment, socioeconomic status, and the like. Developing definitions requires the involvement of scientists who are working with the data, familiar with the context for the data, and have a stake in how the data are used.

- *Identification of disclosure risk.* Data have to be published in a way that avoids disclosure of personal identity. To illustrate the point, he shared a plausible example (see Box 2-2) of what might have occurred with a dataset on mortality data. To protect against disclosure, INDEPTH anonymizes the data in their history data. The micro data for the cause of death are also anonymized but using a different set of random identifiers so that the data cannot be routinely linked. An identity map is available for restricted access to the data when researchers with a legitimate reason and with institutional support or backing have a justification for why they need to be able to link the data. The data are still anonymized but are linkable through this identity map.
- *Value proposition*. A significant challenge is to make a value proposition to the 50 sites as to why they should "put a lot of effort and valuable time into supporting this process within INDEPTH," Herbst said. While the initiatives by the PHRDF have helped make a case, he said sufficient recognition is lacking for the intellectual contribution in making the research data available. He posed the question of how to address this given the struggles faced by overburdened data managers, scientists, and leaders of research facilities within Africa.

BOX 2-2
An Example of Disclosure Risk

Kobus Herbst used the following scenario to illustrate how data can disclose personal identity:

> A first-year computer science student at the University of Zululand is from one of the homesteads within the Africa Center's demographic surveillance site. He knows about the Africa Center and the INDEPTH Data Repository. He self-registers on the INDEPTH Data Repository, and then downloads the Africa Center's data in the repository including the individual-level mortality data with date of birth and date of death. He remembers a friend who died and wonders if he can find her cause of death from this dataset. He knows approximately her date of birth and almost exactly the date she died since it was a major event in his community. It turns out by cross-tabulating month of birth and month of death, some 93 percent of the cells would have fewer than three individuals in them. So it was a simple matter to find the record with the cause of death of this person.

Herbst highlighted solutions INDEPTH developed for dealing with the challenges and for creating capacity. He suggested theme-based joint data analysis workshops are an efficient way to harmonize data and to generate multicenter and center-specific publications with quality-assured and documented datasets. When a special issue of a journal is published—for example, a recent *Global Health Action* with approximately 21 publications—the dataset is also published in the INDEPTH repository.

He pointed to the master's degree in research data management at the University of Witwatersrand as an example of beginning to build careers in research data management. To help address technical capacity, INDEPTH developed a "Centre in a Box," a portable information appliance with a standard open-source environment for any database. The iSHARE (**I**NDEPTH **SH**aring and **A**ccess **Re**pository) project[9] has a help desk to which sites can call for support. It can also be used at data analysis workshops to collaborate with other centers in a controlled environment to access and document data.

Finally, INDEPTH's data curation workshops train managers at participating centers to use the Centre in a Box and the associated toolset. Each participating center receives a fully configured Centre in a Box and data extracted from center databases in a common intermediate form; there are common and standard procedures to process the data further and calculate quality metrics and then document the data using Data Documentation Initiative metadata.

Discussion

The plenary discussion in this session touched on several of the issues raised by Herbst and Carr, including opportunities for cost sharing, the notion of fair trade, data ownership, and ethics.

Cost of Data

It was suggested by a participant that cost sharing would advance development of HDSS data. However, Herbst expressed doubt that a scheme for charging for data and using those funds for longitudinal data collection efforts would be sustainable in the era of open access. Cost sharing, however, gains visibility for a project/site and thus attracts more funding to sustain the longitudinal data center, he said. There is also

[9]INDEPTH established the iSHARE repository to share its data widely and freely on the Internet. The repository has data from three Asian member centers (APHRC–India; Kanchanaburi, Thailand; and Wosera, Papua, New Guinea) and three African member centers (Agincourt, South Africa; Dikgale, South Africa; and Magu, Tanzania).

benefit from having the data used in the North because the findings and questions contained in the analysis help to develop good proposals and receive additional funding.

The cost of collecting and maintaining HDSS data is a concern, particularly when there is limited capacity to analyze the data to influence local policy and demonstrate its value. Funding constraints have further reduced the frequency of data surveillance at some HDSS sites. The goal, a participant suggested, should remain to develop the capacity for African researchers to identify the questions that are relevant to health in Africa, along with the capacity to answer those questions using African data.

Fair Trade

A participant commented that while the discussion should not be about buying and selling data, there should be a "trade" that recognizes the currency that the North and the South have to offer. Another participant suggested the overcapacity of analytic abilities in the North should be used to build capacity in the South. While there is a polarization in skills on data analysis between the North and the South, another participant cautioned against a lens that assumes analysis occurs only in the North.

Another participant pointed to a problem in the context of intellectual property.

A key issue is data ownership. Intellectual ownership of the data implies the ability to be recognized for collecting, preparing, and sharing those data. Fairness calls for the "foot soldier" who collects the data to get credit for the work, one participant stated.

Ethical Imperative

A participant questioned the articulation of the ethical imperative as focused only on the research participant and the potential harm of secondary analysis. He referred to lessons from systematic review methodology where primary researchers wonder if their research is being used, and if so, in a proper, relevant, and pertinent way. If there were a standard on how to conduct secondary analysis, such as having a protocol submitted to peer review, secondary use would provide answers to new questions, the participant said.

iSHARE

A participant asked about the discrepancy between the number of INDEPTH sites (n = 50) and the number that have data on the iSHARE

platform (n = 29). It was explained that the discrepancy is related to capacity, with the goal to expand to more sites.

Data Curation and Management

A participant urged more training at the bachelor's or honors levels, as well as fellowships and other forms of support for candidates interested in data curation. There are also opportunities to share expertise across centers or partners with varying levels of expertise, the participant observed.

Data Analysis Capacity

A participant conveyed a concern about ensuring adequate salaries for dedicated staff to conduct data analysis in African centers. There is sometimes little time after finishing data collection for a project and publishing a report to conduct additional analyses, since the focus shifts to looking for additional research projects. It was pointed out that the data management plans emphasized in funder statements provide a great opportunity, as they represent a commitment by funders to provide resources to manage the data resource after release of the initial reports.

Community Engagement

In response to a question about lessons from INDEPTH around community expectations, knowledge and understanding, and benefit sharing, Herbst described several initiatives, including cooperatives that involve data subjects and an experiment at the Africa Center around "data everywhere," in which one of the components is an interactive environment where community members can explore the data available on the community itself. Herbst commented that it should be expected that communities will have access to their data, which should be considered in planning.

Other Types of Data

According to a participant, sharing longitudinal data as collected by INDEPTH may be easier than some other types of data since the sites collect the data in the same way. Herbst commented that there are limits to which data can be harmonized, and that involvement of scientists with local knowledge and insight into the data is key to harmonization.

Data on Sensitive Issues

Some data that are collected are particularly sensitive, such as data on sexual behaviors and violence exposure, a participant said. People often agree to provide sensitive data because they trust the researcher and trust the researcher is not going to use the data in a way that would be harmful toward them. However, there is a concern that the trust may not carry over when the data are shared. The challenge is to preserve the original consent when the data are shared. Carr agreed that preserving consent is very much at the heart of the debate. It is crucial to respect the consent terms under which data were provided, particularly when talking about historical data.

Qualitative Research Data

Protecting the rights of individuals who give qualitative interviews is also a concern. While the INDEPTH network does not collect qualitative data, the global debate on open access to data includes qualitative research. In one participant's view, the risk is sufficiently great that qualitative data should be removed from the open-access discussion. These data, Herbst agreed, are inappropriate for direct access on the Internet and call for more protected, controlled data enclaves as an alternative.

U.S. Data Sharing

These issues can be covered, it was suggested, in policies that require agencies to have plans for sharing data in an environment that both protects proprietary rights and assures confidentiality and privacy. A participant noted that all U.S. science agencies are expected to have such plans. The plans do not provide open access to everything; indeed, some data are restricted because of the risk of re-identification if combined with other datasets. These concerns mitigate against a blanket approach to sharing all data, the participant noted.

3

Establishing Equitable Terms for Data Sharing

Steven Tollman, director of the South African Medical Research Council's research group devoted to rural health and head of the Health and Population Division at the School of Public Health at the University of the Witwatersrand, Johannesburg, opened his keynote talk by commenting that equity and fairness issues tied to communities deserve attention, along with equity and fairness issues related to scientists.

DATA SHARING FOR GLOBAL HEALTH

Discussing the severity of health concerns in Africa, Tollman commented that the response to the complex environment that leads to adverse health outcomes requires more than one institution or research group. Collaborations around data to serve global health reflect a vision of a common world with common problems. Harnessing the most effective collaborative efforts can yield shared returns, particularly to the poorer communities and societies, he said. However, he commented, very few population health-oriented programs in the "so-called Global South" label themselves global health.

As a result, he said, there may be an underlying fear among those in the Global South. As an analogy, he said mineral and raw materials may be extracted without benefit to a community. Similarly, the Global South may be concerned that data will be exploited and extracted, or taken from a community, without a concurrent benefit to the collectors, the communities, or the data. He observed a divergence in collaborations may reinforce

this concern—researchers in low- and middle-income countries (LMICs) are largely data gatherers while colleagues in higher-income countries (HICs) are primarily the analysts and those who add value.

At the same time, the growing sophistication of field-based research (e.g., acceleration of biomeasures, measures of physical and cognitive function, and more sophisticated approaches to analyzing socioeconomic measures) requires expanded efforts to ensure quality. African institutions lack the needed technical capacity to extract, document, archive, and share the data as well as the methodological capacity needed to analyze complex data, putting them at a disadvantage, he said. Resolving this will require long-term investments.

In Tollman's view, science funding represents an investment that should lead to data that can be shared to produce "new analyses, new discovery, and applications to policy and programs," but legitimate concerns among low- and middle-income scientists, particularly African scientists, need to be considered. African scientists "may end up somewhat on the margins" and may feel that the "host communities do not benefit commensurately." This creates "real imperatives for aware and shared leadership. It's scientific leadership. It's funding leadership, and forms of institutional leadership," he said.

Tollman outlined three needs to address these concerns: (1) aware, shared leadership; (2) assessment of the "flow of benefits"; and (3) recognition of capacities required, both in the form of hardware (e.g., equipment) and software (e.g., skills and techniques).

Effective Collaboration

Tollman discussed an unusual partnership between Wits University and the African Population and Health Research Center (APHRC) in Africa and the University of Colorado and Brown University in the United States. In this collaboration supported by the Hewlett Foundation, the partners met annually to reconnect the partnership, which aims to take a systemic approach to promote research, training, and administration across all four institutions. Common investments from the Hewlett Foundation helped with institutional development, joint grants generation, joint research, linked capacity building, improved administration, and collaborative opportunities for students and staff.

Tollman emphasized that leadership of collaborations can take many forms, but they share several important elements:

- A start-up phase to ensure high-level commitment and lay a foundation.

- Strong, transparent, effective leadership, including possibly co-leadership by North-South partners, or African leadership where this type of leadership is a principle of the collaboration (e.g., the Human Heredity and Health Initiative).
- A clear governance structure that addresses roles and budgeting, and has an effective secretariat.
- Strong anchoring institutions.
- Explicit shared goals with active participation and a fair flow of benefits, or mutuality, involving both written agreements and interpersonal exchanges.
- Interpersonal relationships.
- Dedicated resources for collaboration.

Tollman observed that funding partners can set the terms that contribute to structure and create an enabling environment for collaboration. As he said, "If I remember one thing from anatomy, it's that structure follows function. So it is fundamental to know what the function is, what the purpose is, and then to ensure that the structure in the enabling environment supports that."

In response to a participant's question about his experience with the role of supporting institutions in building effective partnerships, Tollman emphasized the vital importance of administrative support to create an "effective administrative organizational platform"; he also said funders can support and sometimes lead the leaders to develop mutually rewarding arrangements, particularly in institutions where there is the potential for leadership.

Outputs and Metrics

One incentive for data sharing is tied to outputs and metrics of research. The traditional outputs of research that academics are judged by, Tollman said, are publications and graduates. Publications are judged for quantity and quality (e.g., journal impact factor, citation index). Graduates are judged by their number, level (master's, PhD, post-docs), and fellowships received, which are all relatively easy to measure. More complicated to measure are the emerging, less traditional metrics that result from data sharing (see Box 3-1).

As datasets are made available, a range of issues emerges, from the ability to make the datasets publicly accessible, to the demand for the data, to making some assessment of their actual use, as well as whether, how, and what is expected regarding attribution. Other metrics that he acknowledged are more difficult to measure include whether the shared

BOX 3-1
Metrics of Data Sharing

- Datasets
- Tailored/public access
- Number available
- Demand (hits and downloads)
- Actual use: Documenting second line output
- Attribution
- Secondary data use: Scientific or policy
- Policy and program impact
- Nature, extent, and "level" of impact
- Returns to study community
- Public/community engagement

SOURCE: Steven Tollman's presentation.

data are used in science or policy, their impact, and whether the community is engaged with the work. Tollman posed the question if impact and engagement are valid metrics of achievement, and, if so, challenged the group on how to include them in systems of reward and recognition. Assessment panels, he said, are less familiar with and convinced of these metrics' value.

Tollman argued that the cycle of recognition and reward should extend consideration of outputs that gain recognition to include those above and that they should be used to inform assessments and influence rewards, such as promotions and grants. In response to an audience question about measuring impact and how measures would be used, Tollman responded that recognition, promotion, and status should derive from impact measures that include datasets as an impact. Another participant pointed out the creation of datasets is a genuine public good, as they can be used repeatedly and never get exhausted. The impact question is important not only because of the desire to be better at explaining the benefits of sharing data, but also answering questions like, What has actually changed? How has behavior changed? How has public health improved as a result?

DATA SHARING AS AN ELEMENT OF A DATA CONTINUUM

In the discussion period that followed Tollman's presentation, Catherine Kyobutungi, director of research at APHRC and a member of the board of trustees of the INDEPTH network, presented an analogy of data sharing to a hippopotamus. In a hippo's natural habitat, most people only ever get the opportunity to see the animal's eyes and the top of its head, with the rest of its body submerged in water unless conditions are right. "This is what we need to do in the context of data sharing—create the right conditions," she said. She argued that creating the right conditions requires consideration of fundamental issues tied to data use and analysis, and broadening the discussion beyond those at the workshop and beyond Africa, to public decision makers, different cadres of people, students, lecturers, and university administrators.

Kyobutungi said she framed her comments from the perspective of an African researcher in a nongovernmental African research institution. From this perspective, she said, it is important to think of data sharing in the context of funding cycles. At the beginning of the year, her institution has "no committed funds for anything," she said. Toward the end of the year, the focus of the board is how to continue to cover the existing staff the following year. Having conversations around data sharing is not a priority.

In addition, funding for projects versus funding for core activities is a huge issue, she said. Over the past 13 years at her institution, core support has dramatically decreased as a proportion of total funding. In 2001, core support was more than 50 percent of the annual budget. Project support increased dramatically, but without a parallel increase in core support, which is now only 7 percent of the annual budget. Thus, 95 percent of her staff costs are associated with delivering on projects and on sustaining new projects. Expectations for fundraising and project management are huge. At her institution, data sharing is a core support function, so it is considered a luxury, although people want to share data.

APHRC's first data access and sharing policy was drafted in 2007, and the executive director led the INDEPTH data sharing and access policy process. They have a micro data portal that is part of their core funding to make 28 datasets publicly available as part of the INDEPTH Sharing and Access Repository (iSHARE) project. She said the cost is viewed as an investment and an opportunity to inform future efforts. They also view data sharing as a continuous quality improvement process. The iSHARE experience made them realize that data sharing raises issues about quality; thus, APHRC also sees data sharing as a way to improve capacity and improve data.

Returning to her hippopotamus analogy, the real hippo emerges when looking at data generated and collected in Africa, she said. Some data are lost at processing and analysis, an even smaller percentage is available for re-analysis, and a very small percentage is appropriate for sharing. To illustrate how data are lost, she cited sample sizes that are too small or losses at the re-analyze stage. "The biggest data loss is at the re-analyze stage," she said. "Once the project is closed, it's closed. So there are mountains of data sitting even in our own institutions." Shared data are a miniscule portion of the data collected in Africa. "If we need a pipeline for sharing, we cannot have that pipeline if we are losing data across the continuum," she said.

A Uniquely African Issue?

Kyobutungi posed the question of whether data sharing is a uniquely African issue and argued that to a large extent, it is. "Why is it possible for a student in the North to request data from us at APHRC or from iSHARE, but not the other way around?" she asked.

She gave two possible answers, the first having to do with the source of the data. "If I went to Harvard [and] checked the Harvard data-sharing platform . . . maybe 80 percent of it is about America. It is more likely that somebody from Harvard wants to analyze data from Kenya than somebody from Kenya wants to analyze data from America." U.S. data are available, she said, but are unlikely to answer questions of interest to African researchers or institutions.[1]

The second reason has to do with greater access to data and thus deeper exploitation of databases in the Northern Hemisphere, she submitted. She pointed out that it is highly likely that a student from the South would find that an inquiry to the Harvard data-sharing platform would already be answered. The number of scholars, for example at the master student level, with access to data from the North means that the data are deeply exploited. On the other hand, the South has many unanswered questions, she said. She suggested that the Northern students could have 15 potential questions for every 3 that have been answered by African researchers. So in a way, she said, this is a uniquely African issue. "It's not

[1]A participant made the point that any large research project in the United States is required to address in the application for research funding a data-sharing plan and that there are data archives in the United States. Certain well-resourced large studies make data available themselves through study websites, and most studies, including studies that have terminated, have accessible data; for example, the University of Michigan's ICPSR data archive. There are levels of data access, with some available through public release, some with restricted use that requires an approved data use agreement, and others that are considered sensitive and may be accessed through data enclaves.

that Africans are resistant to data sharing, but circumstances and magnitudes are different," Kyobutungi said.

Northern institutions can be fast adopters of data sharing because they have data-sharing practices, platforms, policies, and metrics in place, she said. African institutions operate in a different resource environment. The environment for data access is not the same for APHRC, with its small amount of core support, as it is for Harvard or the London School of Hygiene, she commented.

Capacity

Kyobutungi argued that "if we had the resources, perhaps we wouldn't be so afraid [to share data]," noting if African institutions were able to use their data more, it would not matter. She argued that capacity is needed to "generate good data, capacity to process it, [and] capacity to use it" and that sharing would then follow as a logical outcome. Building individual capacity is not enough, she said. The environment in which individuals operate has to be enabling. The conversation needs to broaden beyond workshop participants and may have to start with the basics about data quality, sharing, and the potential benefits given that data are available. Institutions need to have the right policies and the right guidelines and procedures for research contracting, citing the Council on Health Research for Development (COHRED) as an example. And, she continued, domestic funders have to be brought into the conversation since "a lot of our universities, public universities, are funded domestically." Addressing capacity challenges, she said, is long term and must consider the skills needed and people committed to contributing to the whole cycle of data generation. Kyobutungi suggested that funders consider requiring a capacity-building plan to accompany the currently required data management plan. "If you are demanding data sharing, data sharing has to go with capacity building. . . . We need to embed capacity-building initiatives within the small grants that we are working with," she said. Research institutions need more core support for the activities to enable data sharing.

Kyobutungi made the point that while APHRC has a small micro data portal, there could be benefit over the longer term of "regional initiatives where there is one institution that does archiving." The regional archive could respond to requests, rather than every institution with data have an archiving and sharing function in perpetuity. She pointed to many existing African health networks as those that should be nurtured.

She urged broadening the conversation beyond the current workshop to public decision makers and other cadres of researchers. This requires looking at the whole continuum of data production—data generation,

curation, management, and analysis—and not viewing data sharing as the primary outcome of the conversation. Considering inputs at the front end will facilitate data sharing as an outcome, she concluded.

H3AFRICA CONSORTIUM: DATA SHARING, ACCESS, AND RELEASE POLICY

Michele Ramsay, director of the Sydney Brenner Institute for Molecular Bioscience and a professor in the Division of Human Genetics, the National Health Laboratory Service, and the University of Wits, spoke in her capacity as principal investigator in the Human Health and Heredity in Africa (H3Africa) Initiative.[2] H3Africa started in 2010, with funding from the Wellcome Trust and the National Institutes of Health (NIH), to enhance genomic research on African populations conducted in Africa rather than relying on work done in Europe and North America. The project is a partnership focused on capacity building, with the goal of enabling data producers to also become primary data users and analysts. There are currently ten collaborative centers, seven smaller research projects, three ethics project (with an additional three pending), three biorepositories, and a pan-African bioinformatics network.

The project went through multiple phases before it was ready to share data. It did not begin with an ethics component, she said, but this is now an important part of the initiative with discussion about informed consent, broad consent, and sharing not only data but also biological samples. The project's three biorepositories make the biological resource available for future data generation. If the resource is used to generate new data, the new data have to come back to H3Africa to share with all those involved. The pan-African bioinformatics network, called H3ABioNet, focuses on data management, storage, and analysis. H3Africa promotes fairness by ensuring that the lead institutions are based on the African continent, reflecting a commitment to build capacity. There are also collaborations with institutions in multiple countries in Africa as well as partners outside the continent.

The primary goal of the consortium is to derive the greatest possible benefit from the data generated. H3Africa operates by having a series of working groups. One group has focused on data harmonization. She reported that continued work is needed to ensure the data are equivalent for analysis. Another working group is focused on developing a policy on

[2]For additional information about H3Africa, see http://h3africa.org/ [August 2015]. Also see H3Africa Consortium et al. (2014). Research capacity. Enabling the genomic revolution in Africa. *Science, 344*, 1346–1348, at http://www.sciencemag.org/content/344/6190/1346.long [August 2015].

data-sharing access and release and includes representation from all the H3Africa projects and input from the funders.

She said the principles developed for H3Africa data sharing include

- maximizing the availability of research data, in a timely and responsible manner;
- protecting the rights and privacy of human subjects who participated in research studies;
- recognizing the scientific contribution of researchers who generated the data;
- considering the nature and ethical aspects of proposed research while ensuring the timely release of data; and
- promoting deposition of genomic data in existing community data repositories whenever possible.

One of the largest challenges for H3Africa has been to engage with African ethics review boards since few of them are familiar with the concept of broad consent, particularly for biological samples. Ramsay noted a possible fear of sharing because of a concern about stigmatization or harm, which has prompted discussion of the benefit of sharing.

Ramsay said the nature of the data—whole genomes or genome-wide genotyping data—made H3Africa decide to leverage existing community data repositories in a format that is the same as other formats so data can be retrieved, analyzed, and compared, rather than build new repositories. They also discussed mirroring the databases on the African continent to make them more accessible to African researchers. Their data will include phenotype (e.g., demographic, health, and disease) as well as genetic data, and will enable analyses of the connections between genetic variation and phenotype.

H3Africa has developed a policy in terms of data sharing, access, and release that builds on the principles discussed earlier and that aims to strike a balance of ensuring safeguards to protect the participants and the people who generate the data, while maximizing the ability of investigators to advance their research and then promote wider research.[3]

H3Africa Data-Sharing Timeline

H3Africa has a data-sharing access and release policy that articulates managed access rather than full open access via a complex timeline (see Figure 3-1), Ramsay explained. It allows time for the data generators to analyze the data before the data can be widely analyzed and accessed

[3]See http://h3africa.org/about/ethics-and-governance [August 2015].

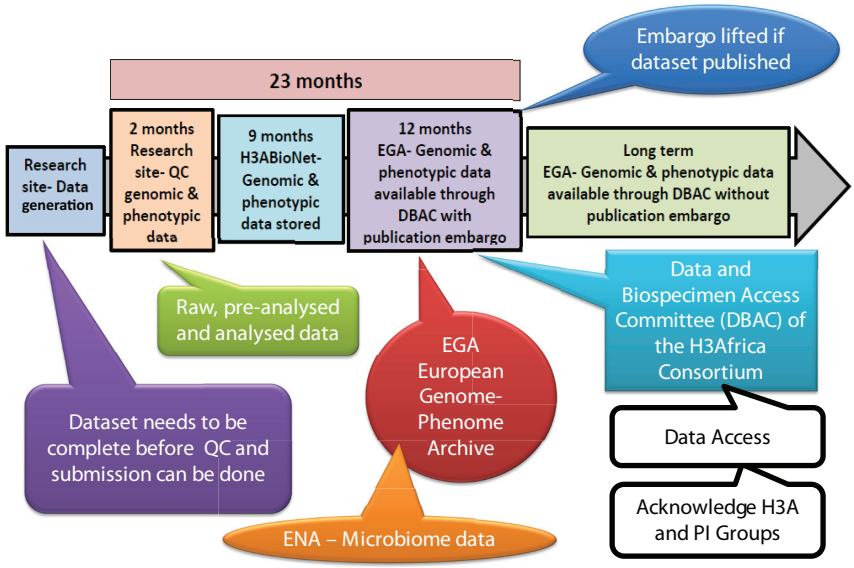

FIGURE 3-1 H3 data-sharing timeline.
SOURCE: H3Africa Data Sharing, Access & Release Policy. See http://www.h3africa.org/images/DataSARWG_folders/FinalDocsDSAR/H3Africa%20Consortium%20Data%20Access%20%20Release%20Policy%20Aug%202014.pdf [August 2015].

by others. It builds on the policies and guidelines of the NIH and the Wellcome Trust in terms of sharing of data, but it has been tweaked to accommodate a sense of fairness for this project on the African continent.

Data and Biospecimen Access Committee

H3Africa is putting in place a combined Data and Biospecimen Access Committee (DBAC) so that requests for data and biospecimens are considered jointly and, according to Ramsay, will be constituted in such a way that it promotes fairness. The DBAC will represent different disciplines and include a layperson. The majority of members will be Africans. The committee will develop policies for internal and external access and will review applications to ensure compliance with what the ethics review boards originally approved and the consent given by the participants. Every project will be slightly different, Ramsay noted, and it will be a challenging process in that it will require review of the original consent and assessing the benefit to be gained from the research in each instance.

The DBAC will be doing scientific review, but they will want some evidence of previous scientific review of the proposals. Researchers who are approved to access the data will be asked to prepare a summary of their study to post online so other people can see what kinds of analyses are being done. They will also sign a data-access agreement that will delineate the research to be done.

Researchers who access the data will have to ensure that they will provide confidentiality, that they will not try to identify the individuals, that they will keep the data secure, that they will not share the data with others who were not named on the application, that there will be legal compliance, and that they will acknowledge the data generators and the bigger H3Africa Consortium project. They will also be asked to provide an annual report to the H3Africa coordinating centers.

A participant raised a question of whether secondary analyses should be required to be reviewed by ethics review boards as well, and whether this might mitigate concerns. The discussion indicated that this policy varies across universities. The DBAC for H3Africa will put caveats on the review and look for evidence of scientific grounding but require projects to go back to ethics review boards.

Long-Term Data Storage

Ramsay explained that long-term data storage will be in the European Genome-Phenome Archive (EGA), which mitigates the issue of needing capacity to store and analyze and release the data, because EGA will do that work. EGA safeguards include a double de-identification of samples so that it is very difficult to link them back to the origin and an encryption process, Ramsay said. Non-human data (e.g., microorganism data or microbiome data) will go into the European Nucleotide Archive.

Fairness

Ramsay pointed out the policy has not yet been tested since H3Africa is now in only the second year of a 5-year project, with the first datasets expected this year. She said that everybody recognized the added value through sharing and is comfortable now with this issue of managed access. There is a slightly longer timeline for the data generators to use the data than might normally be the case for other projects, but she said there is a willingness to share. She also pointed out that H3Africa tries to encourage people to collaborate so that they can build capacity for data analysis, rather than "just being independent users." Ramsay said she thinks it important to analyze data in context, something that H3Africa allows.

Sustainability

A participant raised the question of sustainability since the initiative is funded outside Africa. Ramsay responded that sustainability is an important concept. The consortium is working to ensure that governments help sustain the initiative by supporting it as an important area of research. She reported the South African Department of Science and Technology is involved as a co-investigator. As part of the H3Africa Consortium, there is also an outreach program to garner wider support from governments since the project is building an incredible bioresource and data resource that really will be used for future research, she added.

Cultural Differences

A participant observed that H3Africa includes a number of African countries and asked whether this has raised harmonization or other issues given cultural and other differences. Ramsay responded with an example of DNA sharing, which, like data sharing, has a history of concern in some countries when DNA is proposed to be sent outside their borders. Resolving the DNA sharing issue has meant multiple conversations to gain an understanding of the fears and explain the benefits.

FAIR RESEARCH CONTRACTING

Jacintha Toohey, policy project adviser of COHRED, presented COHRED's Fair Research Contracting (FRC) Initiative and how it contributes to equitable data sharing.[4] According to Toohey, "COHRED contributes to the global health and development arena in a unique manner by enabling the growth of research and innovation systems in low and middle-income countries." They believe that achieving and sustaining global health research is dependent on the capacity of LMICs to use science and innovation to solve their priority health and development problems, both on their own and in partnerships with HICs, as well as within their own arena of researchers, innovators, and institutions.

Toohey noted the Canadian Coalition for Global Health Research has called for acknowledgment that "the persistence of inequitable LMIC-HIC research partnerships be acknowledged." In 1992, the Hesperian Foundation published the book *Where There Is No Doctor: A Village Health Care Handbook* aimed at putting essential health knowledge in the hands of

[4]For additional information about COHRED and its work in Africa, see http://africa.cohred.org/ [August 2015].

LMICs without access to health care.[5] The FRC wants to achieve a similar goal for health research contracting where there is no lawyer, she said. To do that, they have developed tools for researchers in LMICs who do not have legal capacity or frameworks to negotiate global health research partnerships.

In 2006, the International Centre for Diarrhoeal Disease Research, Bangladesh, raised the issue of inequitable research-contracting practices with the World Health Advisory Committee on Health Research and then took up the issue of international collaboration on equitable research contracts with COHRED. In 2009, David Sack and colleagues published "Improving International Research Contracting";[6] and in 2011, a think tank was convened to identify the issues problematic in research contracting practices. In 2012, a forum was held in Cape Town to identify these issues further and discuss how to implement solutions. Experts met again at the Bellagio Centre in Italy, where the concept was established of moving toward fair contracts and contracting in research for health when there is no intellectual property lawyer.

As Toohey explained, the FRC Initiative believes that HICs are called to engage in good partnerships, which means a move toward leveling the playing field in global health research, strengthening of capacity in LMICs to engage in better partnerships, and reducing LMIC dependence on goodwill in HICs. In response, the FRC has identified best practices for research contracting and developed tools that are then placed in the hands of institutions, as well as governments, with limited legal capacity and also where there is no legislative framework.

Toohey emphasized COHRED's view that good practice in research must result in equitable partnerships and include not just a code of conduct for researchers, but also an approach that considers the contractual and negotiation process as important as the research protocol. COHRED is currently developing the COHRED Fairness Index (CFI), which will foster accountability and transparency in research collaborations with LMICs.

The index is in an early stage of development, but one component ready for implementation is the FRC tool, which can guide discussions when entering into research collaborations and partnerships or when negotiating partnerships. She emphasized not every partner is the same in a negotiation and that striving for equity does not mean striving to

[5]Werner, D., Thuman, C., and Maxwell, J. (2013). *Where There Is No Doctor: A Village Health Care Handbook, Revised Edition*. Berkeley, CA: Hesperian Health Guides.

[6]Sack, D., Brooks, V., Behan, M., Cravioto, A., Kennedy, A., Ijsselmuiden, C., and Sewankambo, N. (2009). Improving international research contracting. *Bulletin of the World Health Organization, 87*, 487–489. See http://www.who.int/bulletin/volumes/87/7/08-058099/en/# [August 2015].

be the same. Instead, negotiations should involve frank and transparent discussion about how partners can expect to contribute and benefit from collaborations based on each other's capacities and resources.

The FRC identified six key issues that when properly considered by both partners can lead to substantially improved outcomes for LMICs and their institutions, according to Toohey:

1. *Strategies for negotiation*. LMICs should never sign a contract without the opportunity to provide input into the partnership. Each partner should hear the other's motivations, priorities, and expectations of the outcomes. Different types of partnerships will raise different types of contractual issues.
2. *Compensation for indirect costs*. Contracts should foster and promote a full costing model.
3. *Capacity building and technology transfer*. Partnerships should include a commitment to capacity building.
4. *Ownership of data and samples*. Consideration should be given to how to maximize the benefits of owning the data.
5. *Intellectual property rights*. Joint ownership, including of the research findings that come out of research collaborations, should be explored.
6. *Research contracts in (legislative) context*. FRC is studying where there are legislative frameworks that are missing, and how they can promote fairer research contracting practices in countries with no supportive mechanisms.

Toohey stated that fairness in data sharing is a key component of fair research contracting, as well as a component of CFI. Defining fairness in contracts and then measuring it in a certification system are unresolved issues, but will be an indicator in the fairness index.

Toohey highlighted five key areas related to fairness in data sharing with which the FRC is grappling: (1) allowing sufficient time to analyze data and publish before sharing; (2) contracting to ensure fairness to LMIC partners without research or legal contracting offices; (3) contracting with the researchers instead of the institution, which may deprive the research institution of essential resources and influence; (4) insufficient provision for sharing in post-trial benefits, including related intellectual property rights that go downstream of research projects; and (5) LMICs lacking legislative frameworks to properly deal with research and research outcomes.

The goal of COHRED and FRC is to create an environment where all partners are able to negotiate fairer contracts that will enhance research and innovation for health and bring about global health benefits. Refer-

ence to data sharing or data rights in contracts should mirror the data-sharing policy that partners negotiate.

In response to a question, Toohey said the FRC has three tools available: a fair research contracting guidance booklet, a negotiating booklet, and checklists related to policy and frameworks. FRC has conducted workshops with the Swiss National Science Foundation and the London School of Medicine and is implementing the tools with researchers. The CFI is currently looking at indicators of fairness and how to measure them.

DATA PUBLICATION AND CITATION: THE PUBLISHERS' PERSPECTIVE

Caroline Wilkinson, the open-access relationship manager at Ubiquity Press, a small open-access publisher based in the United Kingdom, gave an overview of issues around data publication and citation from a publishing perspective. She explained that Ubiquity Press "aims to return control of publishing to the research community by providing access to sustainable and affordable publication services. The company takes a comprehensive approach to publishing, viewing any of the outputs of academic research as potentially publishable." In addition to traditional journal and book publishing, Ubiquity Press publishes data papers and is experimenting with publishing software and bio resources, any sort of object, digital or otherwise, associated with research.

Wilkinson shared her view that the focus of the open-data movement is on re-use and reproducibility, rather than just widened access, as was the case with the parallel open-access movement. Many open-access publishers, such as PLOS, are now insisting that the data underlying a paper are made openly available. In most cases, open data are deposited in a data repository. Publishers like Ubiquity are also experimenting with data journals.

According to Wilkinson, the highest-profile example of a data journal is *Nature*'s *Scientific Data*, which launched last year. *Earth Systems Science Data* was one of the first and *GigaScience* is a major big data publisher. Ubiquity Press publishes "metajournals," which provide a publishing platform for data software bio resources. Major publishers, such as Wiley, Sage, and Hindawi, are also experimenting with data journals.

Data papers, with the data creator typically the lead author, incentivize authors to follow good practice in releasing and citing data. The paper acts as a proxy for the dataset itself, she explained. It advertises the work, encourages re-use, promotes collaboration, and provides a measure of impact. It makes citation much easier since the data paper is included in the reference list of research papers for a project and the network for

citation of papers is already in place. Citations can be tracked. Ultimately, data papers provide context and enable others to re-use the data properly.

In Ubiquity metajournals,[7] data papers look very much like traditional articles but are clearly labeled to avoid confusion. The key role of a data paper is to provide context for the data. It includes information about the special and temporal coverage, data collection methods, and rationale for the collection. It also contains very detailed information about the data format, including file types, any quality control measures that were applied to the collection, and, very importantly, where to find the data in the repository and how to access them. The paper includes a very prominent section about the re-use potential for the data as the authors perceive it. Finally, there is a clear statement about how to cite the paper.

According to Wilkinson, citation is vital for data sharing to be effective. It provides a reliable means of retrieval and identification, usually by means of a digital object identifier. It promotes data assistance, and possibly most importantly, it provides a means of giving data creators recognition for their work. In her view, publishers can help to promote this by providing clear guidelines on citation. There is no single way to cite data, but good guidelines are available (e.g., http://www.force11.org [August 2015]). Ubiquity Press's guidelines require that the citation includes information about the data creator, the data publication, the repository, the version of the data, and also its identifiers. For example,

> Alexander NS, Wint W (2013) Data from: Projected population proximity indices (30km) for 2005, 2030 & 2050. Dryad Digital Repository. See http://datadryad.org/resource/doi:10.5061/dryad.12734 [August 2015].

Wilkinson shared that from a publisher's perspective, copyediting is a step where a citation can go awry, so it is important to use this step to reinforce best practices and ensure that journal guidelines are being followed.

The publishing community is working toward having data citations in machine-readable format. Many data re-use scenarios involve very large datasets, possibly from multiple sources. The ability to query them all in tandem or recombine them is very important for data sharing in the future, and, she said, machine readability is a big part of that.

She explained two possible methods for making citations machine-readable—XML or resource description framework (RDF). A very common XML standard is Journal Article Tag Suite (JATS), widely adopted by

[7]For one example, *Open Health Data*, see http://openhealthdata.metajnl.com/ [August 2015].

publishers but designed for journal articles, so it is not optimal for data sharing. She noted initiatives are under way to improve its compatibility with data publication, for example, by introducing terms such as data title version, license type, and what JATS calls "curators," which will be the data creators.

According to Wilkinson, RDF is arguably much better suited to data description and allows information about the relationship between the data and the resulting research. It is not currently widely adopted by publishers, but there are efforts under way to improve the ease of use and wideness of adoption. Wilkinson concluded by saying there is a lot of work to do, both in terms of developing the infrastructure for good citation and engaging with researchers, but publishers are increasingly embracing data publication and providing the infrastructure and network for researchers, including through use of data papers. A participant observed that the increasing integration of data archives, data repositories, journals, and libraries is going to be a powerful force for research in the future.

In response to a question about the relative advantage of data papers over fully documenting data stored in a repository, Wilkinson commented that data papers are "quite a blunt tool" that emerged to provide a bridge because data sharing is such a new concept. Integrating data papers into the existing publication network via a type of academic article is a way to engage researchers and to overcome the problem of attribution and citation. A citation to a paper will be recognized by Web of Science or Scopus, whereas a citation directly tied to a dataset and a repository may not be. Another participant likened data papers to the profile papers now being published for cohorts, often in international journals of epidemiology. Those papers have a similar objective—helping the scientific world understand what the cohort is, so that others can use it effectively.

PLENARY DISCUSSION

Robert Terry, senior strategic and project manager with the World Health Organization, opened the general discussion by commenting that several speakers emphasized the need for support for the whole research process: that generation of quality data is needed to ensure that data are able to be shared. He also observed the call for sustainability funding and core support, the point that sharing data is a form of quality assurance, and that capacity building is needed, including possibly related to data curation and management.

A participant raised a question about the distribution of responsibility to create equitable situations. One panelist suggested that it requires a negotiation of lead institutions to make sure all are in agreement. Another

panelist argued that primary responsibilities lie with funders, who can set terms for research, and with Southern institutions that have to be their own champions, particularly as part of networks. Southern institutions, the panelist said, should diffuse information such as about policies and contracts within their networks, which will make Southern institutions more informed negotiators. Another panelist suggested moving away from a focus on "North-South," since many of the issues discussed are relevant to "North-North" and "South-South" partnerships as well.

Another participant raised the point that more work is needed to demonstrate that data sharing represents a public good, not just something that benefits individual researchers in the form of additional publications. Panelists pointed to the need to demonstrate an impact on population health.

Several participants reinforced the value of thinking of data sharing within a research cycle, not just as data extraction in its own right. Data sharing is a beginning, not an endpoint, a participant said. Another emphasized that through capacity building in the research production process, data sharing will happen naturally.

Another participant raised the issue of sustainable funding for data-management skills. He suggested that research centers create organizational structures that require projects to commit to using common data and database structures. Rather than exporting data in collaborative projects with Northern partners, the storage and manipulation of the data and preparation of the data can be done locally in the research organizations, he said.

BREAKOUT GROUP DISCUSSION: ESTABLISHING EQUITABLE TERMS FOR DATA SHARING

The participants broke into small groups to discuss two points: (1) terms and conditions for data access to ensure an environment of equity and fairness, and (2) incentives that would motivate researchers to share data.

Terms and Conditions for Data Access to Ensure Fairness and Equity

During plenary discussion of the priorities identified by the groups, facilitated by Terry, individual participants identified several potential terms and conditions for fairness and equity in data sharing. The terms and conditions for data sharing should cover the full data cycle, including long-term data sharing from the project, several participants said. The data provider should stipulate any intended use for secondary analysis and outline procedures to ensure that the intended use is a responsible

one; they should recognize and prioritize potential beneficiaries of data sharing—the individual, the community, the organization, the population at large; they should provide a technical platform for sharing data and train on use of that platform; and memorandums of understanding across institutions should specify review procedures, roles, data platforms to be used, and the like. In addition, there should be provision for broad informed consent that recognizes the range of possible future uses to ensure fairness to participants. Finally, there should be provision for feedback to the researchers who collect the data so they have reassurance their work has been used and how it has been used.

Several participants suggested that data sharing would be enhanced if there were incentives for researchers to share data. The incentives could come in the form of (a) designation of a portion of project funding for data sharing, with amounts tailored to the level of institutional capacity with the understanding that institutions that are sharing data for the first time would need more funding to build capacity; (b) investment in development of data repositories that might include African satellite centers of the repositories that exist in the North and over time, creation of centralized resources in Africa; (c) the establishment of protocols for co-authorship by the researcher who collected the data; (d) procedures for institutional recognition of the value of data sharing; and (e) a funding system that supports the work done by data management departments and provides increased funding for a larger number of shared datasets. The terms of these incentives could be spelled out in guidelines for written agreements that specify roles and responsibilities for the researcher who collected data and those doing secondary analysis, including authorship terms and recognition of both the original researcher and the institution.

Terry closed the session by commenting that he saw that many participants agreed that there is a need for an institutional approach and funding to support that approach. He also reflected on the debate on open access to publications and how open access was once "a mountain to climb" but is now commonly accepted. While the field is at the beginning of the road with data sharing, a significant difference is support by journals. He conveyed optimism that data sharing will become the norm if the technical issues around data platforms and the need for more data managers and other capacity issues can be resolved.

4

Exploring the Ethical Imperative for Data Sharing

Michael Parker, professor of bioethics and director of the Ethox Center at the University of Oxford, opened the workshop session dedicated to ethical issues in data sharing.

OVERVIEW OF ETHICAL ISSUES

Parker suggested three ways to think about ethics. First, different approaches to data sharing result in different harms and benefits. Second, what is considered right and wrong is sometimes a separate consideration from the consequences (e.g., sharing data on sexual behavior might benefit science but be considered wrong for other reasons); and professional standards of conduct, or establishing a set of professional ethics for those involved in data sharing, are needed, whether related to collecting data, managing the data in a data center, or managing the sharing of the data itself.

Reasons to Share Data

Parker suggested arguments in favor of data sharing fall into three categories: (1) better science, (2) increased and better health care, and (3) explicit ethical reasons. He highlighted the arguments of each category.

Better Science

Parker noted that discussion at the workshop pointed to data sharing generating more science in a wider range of research and promoting better science. Data sharing may result in better use of science funding, he said, which is especially important in low-income settings. When datasets are unique—that is, it would be impossible to re-create them—they offer particular scientific value, and there are good ethical arguments for trying to use them.

Better Health Care

According to Parker, the better use of data might help to better use health care resources, plan services more effectively, develop more evidence-based interventions, and ultimately lead to better care for patients. He argued that data sharing might therefore be particularly important in low-income settings with high burdens of disease.

Ethical Imperative

Improving health care and generating scientific knowledge create an ethical imperative for the sharing of data. Parker opined that sharing data, if done appropriately, can help to address health inequalities, and therefore creates an obligation to participants who have consented to use the data well and efficiently. He also pointed out ethical implications of *not* using data, raising the question of whether it is more problematic to use samples where the consent is a bit unclear, for example, archive samples, or using additional resources to collect new samples from new participants.

Cautions about Data Sharing

Impacts on Science and Health

Parker acknowledged concerns about being scooped by others who use their data may lead scientists to focus on short-term goals, such as publishing, and be less willing to engage in a more deliberative, strategic approach to their research, which might reap more benefits. It also might undermine scientific capacity in low-income settings, which could have important implications for the future of science. An emphasis on data sharing could provide an incentive for scientists to focus on careers that analyze data, at the expense of generating new data. He noted concerns are also expressed about potential poor-quality secondary data use and the resultant reputational risk for those who produced the data. Data shar-

ing can also lead to opportunity costs as the resources needed for curating and sharing data prevent a focus on areas of new scientific inquiry.

In addition to the scientific impacts that will ultimately affect health, poor-quality analysis of data may impact health quality.

Ethical Cautions

Parker summarized some of the ethical considerations he said he heard raised during the workshop:

- The need to manage *privacy and confidentiality* when data are being shared and datasets can be merged, such that the sharing may generate information that allows people to be identified.
- Concerns about "*moral distance*" or whether the uses of data by those a distance away from where they were collected will take into account the expectations of those who first collected and provided the data in a particular context.
- The possibility of *valid consent*—and if so, is it really possible to achieve valid consent when the future uses of data are unclear?
- Issues related to *social justice*, including stigma and discrimination.
- The potential impact on *public trust* and implication for future research. For example, if data are used inappropriately, such as published in ways that are discriminatory, that might have implications for the trust of communities and the public in the scientific enterprise.
- Issues related to *decision making* and who decides who gets access to data and who does not, and what counts as appropriate involvement in the data-sharing and data-access policy process.

Call for Empirical Research

Parker argued that ethical claims made about data sharing are claims that could and should be tested empirically. "Data sharing is a means to better science, it's not better science in itself," he said. Potential empirical research can range from randomized controlled trials (RCTs) to qualitative research to development of models for data sharing to understanding the benefits of data sharing. Research, he said, could be conducted on questions relating to valid consent, respect, and autonomy; social justice; what it means for research collaborations to be fair; and requirements for public trust and confidence in the scientific enterprise.

For example, research could explore what models of consent work, recognizing that consent by its nature, even in high-income settings, is imperfect. Parker noted valid consent has to encompass information and

understanding, voluntariness, and competence. Achieving these three things is imperfect since the process of obtaining consent is a social phenomenon with real-world people in a complex process. Consent is inevitably less than fully informed since, for example, potential future uses of data are unknowable. Nonetheless, research could help continue to develop an evidence base for models of good practice.

A participant pointed to a substantial evidence base on privacy and confidentiality that could be better used in decision making. Another participant questioned the value of doing RCTs on data sharing since a lot of research has already demonstrated the ability of data sharing to increase knowledge. He suggested instead an understanding of the complexity of data-sharing issues, for example to differentiate between data known to be important today and those that will be important 10 years from now. He commented there is a lag issue that is a hard problem to solve for a data steward.

Data Utility

A participant pointed out tension between data that have broad versus narrow utility. Some data are going to be widely re-used almost immediately. He stated that the case for sharing those data is unequivocal. Less clear is what to share and what to preserve among things that might be very important but have a narrow utility. Another participant shared that funders are struggling with this issue. Funders have a broad policy on data sharing at this stage because it is too hard to know what to keep and what not to keep.

Social Justice

In Parker's view, the inherent imperfection of consent calls for more attention to questions of social justice. Consent alone does not make research ethical, he stated. In addition to valid consent, responsible conduct in data sharing requires protections around discrimination, security standards, and standards of confidentiality and privacy. There should be very good governance and oversight, and the security of the data should be guaranteed, he urged.

Researchers and participants in low-income settings should be able to be confident that broad social justice considerations are also being taken seriously, in his opinion. They should be able to expect that research funders and research institutes are attempting to address global inequalities, that research is socially relevant, and that it is addressing the so-called "10:90 gap"—the view that 10 percent of worldwide resources devoted to health research are put toward health in developing countries,

where over 90 percent of all preventable deaths worldwide occur. How to best address social questions is, in his view, another empirical question.

Public Trust and Social License

Parker said that social license, a concept used in sociology, is relevant to data sharing. Sociologists argue that there is a distinction between the social license given by society to researchers and the mandate claimed by researchers. He said that it is "very important that those two things are close to each other for trust to persist in a research enterprise."

An example in the United Kingdom is historical work in which organs from children who died were retained without the full consent of their parents. The doctors conducting research believed they had a mandate to conduct the research, but it became apparent that there was no social license for that research, and a problem arose. Similar considerations need to be thought about in the context of data sharing, Parker said. While work can be done with communities to help them see the value of data sharing, for continued sustainability, "research needs to be compatible with the reasonable expectations of the relevant stakeholders."

Fair Trade

Fair research collaboration, or fair trade, is an important ethical consideration as "successful science depends upon sustainable scientific collaborations between researchers in low- and high-income countries," Parker said. In interviews he has conducted with scientists in Africa and Southeast Asia, capacity building, fairness and respect, and an opportunity to set scientific agendas and operate at the cutting edge of science are high on their list of requirements for fair collaboration. He suggested an opportunity to develop an evidence base on the difference between good and bad collaborations by developing "high-quality research looking at different ways of managing data sharing."

Data Ownership

A participant raised a question about data ownership, conveying that he views the researcher as the collector and custodian of the data, but not the owner. Parker responded that he does not favor the concept of ownership, although it offers ways to formalize protections and manage exchange of data. Instead, he would rather "focus on the things themselves that are important rather than focusing on ownership." The issue of respect cannot be solved by declaring an owner, he said. Instead, it requires an approach to data sharing that involves setting the stage for

reasonable behaviors at the outset which, in turn, requires reasonable oversight and governance, and a fair exchange.

In response to a participant pointing out that law in certain countries requires a data owner, Parker said there might be a need to have someone accountable by law, and that could be considered ownership if necessary. He gave a UK example where "no one owns a human body when someone has died, but there are all sorts of rules about who has to do what and how it has to be treated."

Another participant noted a shift in the United States away from ownership toward custodianship, a challenging situation especially with national surveillance systems where the states contribute data. The question is not ownership, it is custody, she said: Whoever has the data in their possession has to have responsibility and is the custodian.

In closing, Parker asserted, "We need to think holistically. If we're serious about promoting science rather than promoting data sharing in its own right, then we need to think in a rounded way, and we need to generate evidence about what's the best way of doing that."

In response to questions from several participants about who needs to be involved to move forward, he suggested engagement of as many stakeholders as possible, making sure there are protections in place, and thinking carefully about the justice elements. Ethics committees have an important role to play, he said, but are often less than perfect and do not have adequate resources or training, perhaps especially related to data sharing. Consideration of ethics needs to continue beyond approval by a committee, because "many of the most interesting and challenging ethical issues arise after you've got the ethics approval."

STAKEHOLDER PERSPECTIVES ON DATA SHARING IN LOW- AND MIDDLE-INCOME COUNTRIES: FINDINGS OF A MULTISITE STUDY

Susan Bull, senior researcher in international health research ethics and head of Global Health Reviewers, was the first of several presenters discussing the findings of a multisite study funded by the Wellcome Trust.

Overview of the Study

According to Bull, the study involves the University of Oxford in England, and five low- and middle-income country (LMIC) sites in India, Kenya, South Africa, Thailand, and Vietnam. It is designed to look at the appropriate governance and management for data sharing given the increasing mandate from funders, journals, and other organizations and

given the range of ethics issues that arise with data sharing, particularly in LMIC settings. The aims of the study are to

- understand the perceptions, experiences, and values of key stakeholders in low- and middle-income settings who are involved in data sharing;
- identify principles for development of models of governance of good data-sharing practice that are relevant in these settings; and
- develop resources to support the development of appropriate data-sharing policies and practices in research involving such countries.

The project included five qualitative studies along with a systematic scoping review of the literature. Bull agreed with previous presenters that if data sharing is done properly, it can improve science, and if not, it will hamper science. She observed that the arguments for and concerns about data sharing are two sides of the same coin, which underpins the point that "to achieve the advantages of data sharing, then we really need to look at how we address the concerns arising."

Bull said the study focused on release of individual-level data, not aggregate data, and on studies of public health and clinical research. The majority of respondents suggested that curation would be needed of some datasets. The reasons cited for why this was necessary included the need for safeguards, bona fide access restrictions, privacy, less harmful or poor-quality research, and compliance. Researchers in the study also raised concerns about the commitments they made during consent processes. She emphasized that decisions have to be responsive to the context and the nature of the dataset.

Priorities were identified for policy and process development, she said. From the perspective of prioritization of data sharing, questions posed included: Which data should be shared? Why? What standards might be used to prioritize sharing data? What are appropriate data and metadata standards? More broadly, she said, at issue are the requirements and rewards needed for collection and curation of datasets and data sharing.

According to Bull, the empirical research that went along with the systematic review was the key element of the study and it started with the premise of "flipping points." In this context, flipping points are defined as elements that might make sharing of data that is acceptable to the stakeholders become unacceptable and what is an appropriate response.

Presenters discussed the results of qualitative research focused on understanding the perceptions, experiences, and values of stakeholders

in the study sites, repeating themes that had been raised throughout the workshop.

South Africa

Blessing Salaigwana and Spencer Denny presented stakeholder views of key features of good/ethical data sharing within a South African context based on a multisite study where they sampled purposefully three different research organizations: two mostly involved in biomedical research, and the third mostly in social science research. Denny presented some main findings from the first paper that came out of the project. He reported a mixed level of awareness among their participants of the procedures and policies or issues related to sharing data, but a general consensus that sharing individual-level data at both the local and international level is for the greater good. According to one assistant investigator, ". . . the more that data is made available the more likely it is to lead to scientific impact."

He said the exploration of questions tied to why to share data boiled down to three issues:

1. the recipient of the data who would be conducting the secondary analysis,
2. the value by participants of altruistic action that has global value, and
3. the tension between benevolence and competition.

In the research cycle that they identified, data were described as the lifeblood of the participants' (primarily researchers) work. In the cycle of the participants' careers, data collection leads to exclusive analysis, which leads to publication, which leads to future funding. Free sharing of data does not complement this model. They shared familiar perceived disadvantages to sharing data, such as misuse and loss of recognition for local stakeholders. In addition to the perceived disadvantages of data sharing, they identified several barriers that deter the practice:

- lack of data-sharing precedence in South Africa;
- lack of incentives, with a sense that there are no returns to counter the risks of sharing;
- lack of specialized infrastructure such as data management and curation; and
- insufficient allocation of funding. The typical research grant does not allow for data management and curation activities to happen post data collection.

Among the study participants, there was a sense that the potential harm was greater, given the diminished prospects of benefits after secondary analysis and the geographical detachment between the data source and the end user. The project identified factors needed to minimize potential harms of data sharing, including respect for the interests of the research participants, accurate data management, preservation of professional integrity, and benefit sharing and capacity building.

The participants then discussed the formal ways that data re-use is regulated, including informed consent and the contractual obligations of the principal investigator to the funder. They also suggested participation of scientific review committees. The study did not involve any research ethic board members, but participants saw a potential role for these boards in resolving conflict between parties and protecting the interests of research participants.

Based on their interviews and focus groups, the researchers suggested that ways to facilitate data sharing include alignment of stakeholder interests, funder support for the required infrastructure, a culture of learning from prior examples (e.g., a resource guide), cultivation of collaborations, and ensuring that data-sharing plans and data-management budgets become a standard budget line when applying for research grants and in ethics review.

The research agenda they developed included two components. First, they see more work in terms of national policy developments and evaluation, with the view toward guiding principles in a South African context. Second, they see resource development to support the decision making on the ground by primary researchers and strengthening of community advisory boards.

Kenya

Vicki Marsh and Irene Jao, from the Kenya Medical Research Institute Wellcome Trust Research Programme, presented the findings of the study in Kilifi County, on the Kenyan coast. The study involves about 260,000 people who live in the catchment area of the hospital/research center.

The conversations drew on the participatory skills of the community engagement team and visual aids to help participants understand the concept of data sharing. After setting a basic understanding of the steps in the data-sharing process, participants were asked "What if other researchers would like to access that data?" and "What if those researchers are situated outside of Kilifi, outside of Kenya, outside of Africa?"

Jao reported that data sharing was supported overall, but with caution. Researchers were more strongly positive than community stakeholders. Many of the concerns or challenges identified echoed themes discussed at the workshop and were tied to perceived harms to the participant/

community (e.g., confidentiality, stigmatization, sensitivity of data) or burdens/harms for the researchers (e.g., need for resources for archiving/managing data, potential misuse of data, unfair competition). Sensitivity of data became a particular issue because of concerns about how the data were going to be used. Trust, which Marsh said is a prominent issue in the literature, emerged as an important issue in their discussions.

Jao highlighted three main findings: promoting researchers' and the primary community's interests, respecting autonomy and choice, and ensuring fair governance and accountability.

Balancing of Benefits and Burdens

Rather than the balance of benefits and burdens being thought of as protections, their project focused on promoting interests of the community and of researchers. For primary communities, promotion of interests involved re-use of data in relevant ways to similar populations and through a partnership with the Ministry of Health, which would regulate re-use. For researchers, promotion of interests involved promoting local scientific capacity building, with high value placed on doing this within scientific collaborations.

Autonomy and Consent

Prior individual awareness and agreement were seen as very important to sharing data. Many options were weighed, including sharing data secretly or looking for participants who participated in primary research to re-consent when a secondary request is made, and the broad form of consent. Broad consent became acceptable only as a compromise, not an ideal, and only if linked to fair decision making when data requests were made.

Fair Governance and Trust

Fair governance and trust were discussed in terms of developing policies adapted over time and in terms of how decisions will be made about accessing data in the future. Jao pointed out that issues of fair governance and trust are connected with those of promoting interests, autonomy, and consent. An aspect of fair governance discussed was to have national regulatory frameworks that govern international data sharing.

Community involvement was an important component of governance for their project. Jao pointed out more research is needed on how to go about it, but suggestions included creating awareness to communities

about data sharing, involving them in informed consent processes, involving them in policy development, and possibly involving them in decision making about access to data when requests are made. However, there were concerns that governance structures on their own cannot totally prevent misuse of data.

Jao summarized the lessons from the Kilifi site related to building trust for data sharing as needing to (1) ensure individual prior awareness and agreement, perhaps through broad consent; (2) develop fair governance processes, which include independent and accountable mechanisms, including accountability to communities, promoting local interests for communities and researchers and international data sharing within national frameworks; and (3) promote data sharing within scientific collaborations.

Thailand

Bull reported on the project on the Thai-Burmese border, a permeable border with an informal, vulnerable population that includes a migrant population that often has no legal right to be in Thailand. It involved interviews with senior researchers and junior research staff in Bangkok, and interviews with community advisory board members in the Shoklo Malaria Research Unit. The project did not include interviews with research participants because the ethics committees in Thailand thought the participants were too vulnerable and would be harmed by asking their views on data sharing.

In this site, although participants were generally in favor of data sharing, there was a very broad lack of experience, even among senior researchers, of sharing data outside research collaborations. Reservations were raised about potential harms to patients and communities, to researchers and research groups, and about the availability of resources required for effective data sharing. There were also concerns, similar to those raised at the workshop, about quality control and experience. The value of sharing data within research collaborations was familiar and very welcome, and a real core focus on determining how to get data quality that is appropriate for sharing along with questions about appropriate consent models.

According to Bull, there was not clear consensus in this site about broad consent or specific consent. Instead, there were questions about what appropriate models should be, and the desire for proportionate and fair governance processes that are responsive to the data being shared.

India

As Bull explained, the project in India involved the Society for Nutrition, Education, and Health Action, which works with women and children in informal settlements in Mumbai. Their research interests focus on child nutrition and infant feeding, with programs to address severely malnourished children, maternity care, domestic violence, family planning issues, and safe abortions. They collect empirical health service intervention data.

Like other sites, this study found that participants were generally in favor of data sharing, but most had very limited experience, and it was difficult to find participants with an experience outside of collaborations. The reservations expressed were related to power imbalances and previous exploitation with these populations. Research participants also said research should be responsive to the health needs of the community and were concerned about confidentiality, even more so than consent. They were concerned about good governance, confidentiality, and making research responsive to the context.

Field workers echoed these sentiments, and emphasized their responsibility for maintaining trust with the women served. They expressed concern that building relationships and collecting the data is very hard work, and secondary users could jeopardize that.

More senior researchers expressed concerns about harmful data use in terms of inappropriate secondary analyses, how to manage excess to preserve participants' interests, and issues about ownership, control, and authorship. The primary finding was in the importance, given the unfamiliarity and the complexity of the topic, for demystification and clarification.

The project developed a model to illustrate funder pressures, policy changes, and the inertia tied to practical barriers (see Figure 4-1). While not necessarily an ethical component, a quite strong practical issue, according to Bull, is that collecting the data takes a huge amount of effort and there are not sufficient policies and processes to provide confidence that there is good governance of data sharing. Further, there are skepticisms and concerns about who the accessors are, what the harms might be, and what the effects for researchers of sharing data would be. These concerns lead to suggestions to develop a model data-sharing policy and standard operating procedure, provide resources for quality control, and prepare data for curation. This is more than just providing resources to inform this process.

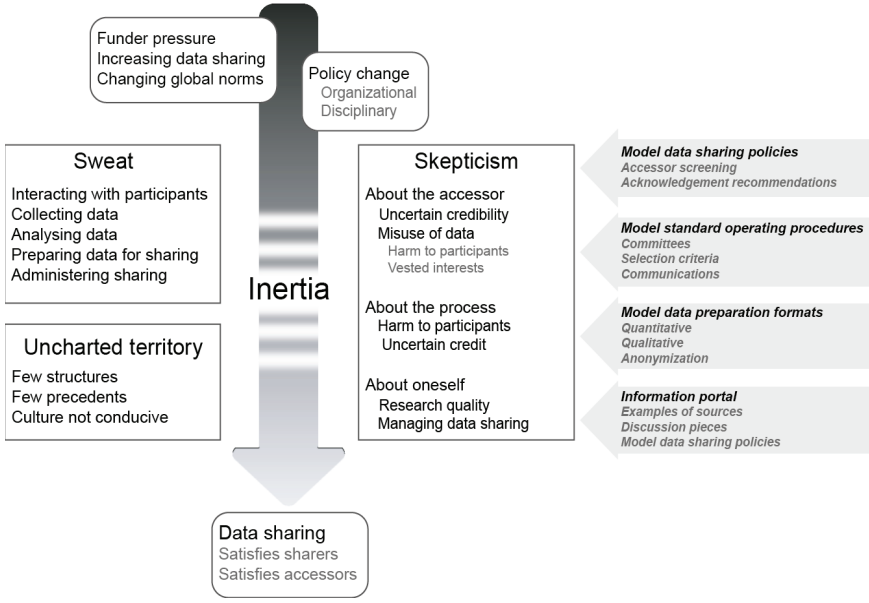

FIGURE 4-1 Model of the data-sharing process from the SNEHA Research Study. SOURCE: Hate, K., Meherally, S., Shah More, N., Jayaraman, A., Bull, S., Parker, M., and Osrin, D. (2015). Sweat, skepticism and uncharted territory: A qualitative study of opinions on data sharing among public health researchers and research participants in Mumbai, India. *Journal of Empirical Research on Human Research Ethics, 10*(3), 239–250.

Vietnam

The study in Vietnam, with the Oxford University Clinical Research Unit, was conducted in Ho Chi Minh City and Hanoi and also in rural areas, which gave a geographical spread. The study, Bull said, aimed to compare opinions from north and south urban and rural settings. Respondents included researchers and ethics committees, but participants had limited experience of data sharing. Data sharing was recognized as valuable in theory but not seen as a priority issue. An unusual finding for the study, not replicated across the sites, was very high explicit levels of trust in researchers and the governance mechanism.

Compared to other settings, Vietnam has a huge amount of governance about many aspects of day-to-day life, including the conduct of research, Bull explained. In Vietnam, in response to these very high levels of trust, there was an acceptance that broad consent was appropriate given the community benefit.

Researchers in turn had a strong sense of responsibility toward patients, as did ethics committees. They felt there needed to be oversight of future research data uses to preserve the interests of researchers and participants. Similar to other sites, experience with sharing data was primarily through collaborative relationships. In addition to collaborations, there was a strong emphasis on authorship being preserved in publications, including secondary analyses. This was not only to give recognition but also to ensure control of the uses, the secondary analyses, and things that are not published that will harm the population.

The project discussed draft principles for governance and policy, and what the priorities might be, which included the following:

- To ensure that the rights and interests of research participants and their community are safeguarded, including preserving privacy, the right to dignity, protection from harm, and appropriate sharing of benefits.
- To protect rights and interests of primary researchers, particularly given the potential inequalities in resources available to support local analysis and publishing.
- To be transparent and accountable.

Bull noted at this site, research participants and junior researchers were most uncomfortable being asked about their opinions. Because there is not a clear national framework and policy environment for these decisions to be made, asking someone to venture an opinion is considered disrespectful. This presents an interesting contextual issue of how to engage with some populations and, given the emphasis on stakeholder views, highlights a context in which soliciting stakeholder views can be disconcerting and perhaps even threatening in some cases.

Findings of the Five Studies

Bull reported three broad themes across all five sites, despite their differences:

1. *Protection of research participants' interests, not just their privacy*. While appropriate consent is needed, there is also a need to minimize stigmatization, and for some datasets, a curated process to promote benefits and minimize harm.
2. *Fairness and reciprocity*. Participants and communities both need to see benefits of research, including research that addresses locally relevant issues and benefits health. There is also an issue of fairness for researchers and institutions in terms of building capacity, benefiting from collaborations, and getting recognition.

3. *Trust and trustworthiness*, including ensuring scrutiny and control of secondary access to data as a mechanism of promoting trust, community engagement, and appropriate levels of stakeholder participation in decision making, ensuring that sensitive data are actively protected, and ensuring data quality.

Bull observed that information from the five sites conveyed that whether data are sensitive goes beyond what the data are (e.g., HIV status). Who uses the data and how they are used were viewed as important considerations in whether data are considered sensitive. As an illustration, she said, "particularly in informal settlements in Mumbai and on the Thai-Burmese border . . . some of the most basic data about economic status can be extremely sensitive depending on who gets their hands on it."

Bull reiterated most data sharing in the sites was done in collaborations. For those researchers involved in collaborations, concerns about data sharing, including concerns about curation and appropriate methods, "fell away"; in fact, they suggested that data sharing is the way research should be conducted. Researchers in collaborations establish trusting relationships, have an opportunity to build capacity, and have the chance to protect participants, she said. She noted these researchers felt that research can be locally responsive and better quality; as a result of educating partners about contextually specific parts of the dataset, the research will be better. However, collaborations are very resource intensive and have the potential to substantially restrict data use, she said.

Developing Resources to Assist in Policies and Practices

Bull said the findings from the project will be published in a special issue of the *Journal of Empirical Research and Health Research Ethics* in July 2015. The project also expects to publish additional papers. The qualitative datasets from the study will most likely be available via the UK Data Service. The project is also developing an online toolkit available through the website for the Global Health Network.[1]

In addition to providing resources, the intent is to provide a site for people to contribute blogs and facilitate discussions. A free online e-learning course will cover how to think about data sharing when developing a protocol, drawing heavily on resources from other sites. The site will include policies and processes, as well as lists of data archives.

In closing, Bull called for input and suggestions from workshop participants about other available resources, as well as needed resources. She

[1]See https://tghn.org/ [August 2015].

said they are trying to make resources available in a collaborative way and "not reinvent the wheel."

Discussion

Levels of Understanding

A participant asked about the level of ignorance about data sharing in the African study and whether the researchers were developing a data-management policy. In response, it was confirmed that there is a mixed level of awareness, or ignorance, with data sharing. Senior researchers had knowledge and experience through work in collaborations. For junior researchers and community stakeholders, the study used hypothetical vignettes to depict the process of data sharing to elicit views. It was further pointed out that the institutional context had an impact on that mixed awareness, and different institutions had different perspectives. In the examples, data management emerged as a key area but not because it was being probed for. In discussion around data sharing, data preservation came up as a key area.

Community Involvement

Participatory methodologies are used to get participants engaged when they really do not understand the process. The study in Kenya drew heavily on participatory methodologies. The researchers also drew on a network of 200 representatives elected by their own communities. They now meet with the network two or three times a year. The group of people is widely representative, but has an atypical understanding of what research is. They tackled the technical aspects of the discussion by beginning with people with whom the researchers had a prior relationship. These participants had basic understanding of some of the topics, even if not about data sharing, but at least about research and what research is for. Likewise, during the focus group discussions, the researchers tried to get from the participants their views about how to involve communities in data-sharing issues.

Trust in Vietnam

A participant asked whether the strong sense of trust of participants in Vietnam reflected a power imbalance in the society. It was noted that Vietnam is a strongly hierarchical society, and that the trust placed in researchers was very strongly felt as a responsibility by researchers.

Thus, trust was conveyed in a very positive way, and there was a feeling that the right thing will be done.

Re-Consenting

A participant asked whether the issue of re-consenting was raised in the sites, suggesting that re-consent is necessary as new tools and connections with datasets are made. Although generalizations are difficult based on the relatively small study size, the researchers got a sense that people would prefer re-consenting if they thought it were feasible. However, other people felt that it was overly burdensome. The idea of broad consent was distinctly a compromise rather than an ideal.

Similar to concerns about re-consenting are concerns about finding the same people. Other sensitivities include returning to a location where the participant died. Because of these challenges, broad consent became a compromise.

Dialogue on Data

A participant commented that the discussion reflected a UK qualitative study called Dialogue on Data, which looked at the public acceptability of using and sharing administrative data. One of the findings of this study was that it is difficult to get across why people were being asked these questions, because the reaction often was "surely you should just be doing this, just get on with it." It was quite apparent that there was a high degree of trust in researchers to conduct this kind of research, which puts researchers, the participant said, in a very privileged position. One finding that came from this research was a distinction made in people's minds between researchers doing research in the public interest or having public benefit in an institutional setting, universities, and research institutes versus research done by commercial organizations where there would be profit gained from the data that people contribute.

Commercial Gain

A participant posed a question of whether it mattered if the research were viewed for commercial gain, an important aspect of public health research. In Vietnam, commercialization was explicitly welcomed because it was seen as the best likelihood to advance health. In three other sites, commercialization was considered to make data use extremely sensitive. Another panelist responded it may depend to a large extent on building understanding of what is at stake.

Concerns about Sharing Across Borders

A participant asked whether trust concerns were different based on with whom the data were going to be shared (e.g., within the university or with another country). In the discussion, it was noted that sharing across borders is complex. There may be explicit trust in an external organization like the World Health Organization, but local use of the data may be more likely to promote uses that are sensitive in addressing local research issues. Another participant shared an example of a PhD student from Ghana who interviewed people in different African countries about export of blood samples. The premise of the project was that people would be worried about the samples going to the North, but the researcher found they worried more about the samples going to other institutions in their own country or to other African countries. A presenter said the junior researchers and community stakeholders were not necessarily against exportation of data but wanted to know how the local community would gain from the data export.

Continuation of Data Sharing

A participant asked whether there was a sense in the communities studied that data sharing should be stopped until concerns were addressed. The presenters replied that one of the overarching findings was that data were not being shared outside collaborations and that there was widespread unfamiliarity with the topic. They emphasized a consultation process designed to explore a range of perspectives, not to achieve consensus. One presenter reported that in follow-up interviews with a few participants, those most stringent in their views about conditions for sharing tended to soften by the time of the follow-up—perhaps, he posited, because it was initially a novel concept and their views changed after talking with others about it. One person further stated that there was a "sense that we need to build stronger and more responsive policies" and "need to do more research." Another presenter commented that participants with data-sharing policies in place tended to identify issues to be resolved, but not state them in a way that suggested data should not be shared.

PROMOTING BEST PRACTICES IN ETHICAL DATA SHARING

Participants broke into small groups to discuss best practices in ethical data sharing in six areas: (1) capacity building, (2) policies and processes for ethical data sharing, (3) action by funders, (4) actions by researchers, (5) further research and evaluation, and (6) trust and confidence.

Reports from the discussions were presented in a plenary session. While there was commonality across the groups in ideas raised, the ideas were not prioritized or synthesized, nor were the implications of strategies for implementing discussed. The ideas posed by the groups are listed below.

Capacity Building

Breakout group participants suggested the following ideas related to best practices in capacity building:

- Include data sharing as part of the research cycle itself; that is, embed data sharing from proposal development through other stages of research.
- Establish data-sharing centers of excellence. One group discussed enhancing and linking networks of research organizations.
- Utilize existing expertise of centers by requiring stronger centers to be supportive of weaker ones.
- Provide training in ethical data sharing, both on-the-job training and formal training, that would start with development of a body of knowledge informed by empirical data.
- Explore how international bodies, such as the Nuffield Council on Bioethics, can be leveraged to develop good standards of practice.
- Establish more institutional review boards or ethical review boards and more training for those boards on how to understand and evaluate data-sharing plans and the implications of data sharing.
- Improve the capacity for data management to enable data sharing on the ground, in research projects, and in research institutions.
- Build capacity for research support as well as research analysis—while the assumption is that institutions in the South need to build analytic capacity, increased capacity is also needed in areas, such as data curation, administration, accounting, IT support, documentation support, and other research support services.
- Build capacity to understand the processes that make data sharable, including documentation, what makes for quality data, standardization, and, possibly, harmonization. Junior researchers, who are the ones who tend to interface with research participants, may be a particular target. Communities and research participants themselves are an additional audience for capacity building around ethics of data sharing.
- Develop information on the costs associated with data sharing.

Trust and Confidence

Breakout group participants suggested the following ideas related to best practices in developing trust and confidence:

- Use community consultation as a way of defining and obtaining a "social license" to use data.
- Develop specific data governance mechanisms.
- Define the key stakeholders whose trust and confidence needs to be built.
- Develop pilot projects for data sharing to establish a framework and build trust.
- Develop policies in data sharing within institutions to build the trust and confidence among the partners.
- Define the ethical challenges or ethical issues related to building trust and confidence, such as transparency.
- Provide information to research participants on the issue of data sharing and ensure transparency about how decisions are made about sharing.
- Make building trust and confidence standard good practice for researchers, perhaps even considered responsible conduct of research.
- Define research as equal to data collection, data analysis, and data sharing; that is, as a fundamental component of research, not an exception.
- Approach communities with an expectation of data sharing, but work with them to develop data-sharing plans by informing the community about who the data are going to be shared with, why, and what will be learned from it, and based on their response, learn, adapt, and continue to move forward on the plan.
- Share data as a way to build trust in the data.
- Consider trust as an issue that involves everyone from funders to researchers to the communities themselves.
- Recognize historical issues that may make trust difficult for some researchers, including the post-colonial approaches to research with a North-South divide.

Further Research and Evaluation

Breakout group participants suggested the following ideas related to further research and evaluation:

- Do a meta-review of data-sharing policies to generate a standard set of templates for ways that people could organize data-sharing policies in various settings.

- Create an evaluation process to evaluate that template and its implementation.
- Conduct research on informed consent about data sharing, especially in settings where research participants have relatively low levels of education, in order to understand really how people understand data sharing on the ground.
- Conduct further research and evaluation on how to establish trust and confidence.
- Approach data sharing as an empirical enterprise, including on issues such as time to release data.

Policies and Processes for Ethical Data Sharing

Breakout group participants suggested the following ideas related to policies and practices:

- As part of the engagement and consent process, consider how much information needs to be provided; for example, how much information is needed if requesting broad consent.
- Develop policies on what and how often to provide feedback to communities from which the research data came, with the goal of frequent feedback to support engagement.
- Create governance committees that include principal investigators to address use and re-use of data in order to preserve trust, taking into consideration the role of research participants.
- Create data-management and data-analysis plans with clear protocols to address fears of sharing data given uncertainty about how they will be used.

Funders

Breakout group participants suggested the following ideas related to funders and ethical data sharing:

- Agree on what, when, how, and for what purpose the data should be made available and funded.
- Check for noncompliance based on the established data-sharing agreement.
- If noncompliant with the data-sharing agreement, take necessary steps to ensure compliance, recognizing that enforcement is difficult given resource and knowledge constraints.
- Consider developing a code of conduct for funders that would define meaningful sharing in order to avoid token data sharing and facilitate provision of usable data for research elsewhere.

- Consider how to create and enable an environment for data sharing.

Researchers

Breakout group participants suggested the following ideas related to researchers and ethical data sharing:

- Engage communities at the start of research when proposals are first being written and ensure transparency throughout, including about what and why the research is being planned.
- Research institutions need to affirmatively state their commitment to data sharing, similar to what funders have done. Ideally, there would be a statement that different institutions would get credit for if they agreed to it as the policy of their institution.
- Increase awareness of the kind of processes needed to ensure no possibility of reverse identification in legacy data.
- Create a norm of data sharing among researchers, and a system that values analysis of shared data similar to primary collection and analysis in order to attract junior researchers.

Summary Comments

The panelists closed the session by highlighting points that stood out for them in the discussion. Katherine Littler of the Wellcome Trust highlighted three areas: (1) the need for centers of excellence; (2) transparency in terms of consent, processes, and decision making; and (3) the potential to connect timelines to capacity building, as illustrated by H3Africa. Michael Parker said data sharing needs to be thought of as a cycle, rather than a one-off exercise, and that models will "need to be thought about, evaluated, and developed over time." He pointed out that sharing will raise issues for consistency when studies involve multiple countries or locations. He also suggested development of a code of conduct or professional guidelines for researchers similar to the code of conduct for doctors that would say "something about not only the kind of things you should do but what kind of person you should be." Bull noted most of the breakout groups focused on trust and confidence, and capacity building for institutional review boards and research ethics committees was a new and recurrent theme with practical implications. She also noted community engagement was consistently mentioned as a needed component of developing trust and confidence and of capacity building, as well as an area where policies and processes are needed.

5

Enabling Data Discoverability, Linkage, and Re-use

Peter Elias, a social scientist affiliated with the United Kingdom Economic and Social Research Council, and an early and still-active member of the Public Health Research Data Forum (PHRDF), opened the next workshop session. He began by commenting that there is no major challenge facing the world's population that does not have social science causes or consequences or both, and that includes public health.

He also noted that it is important to focus on the practicalities and the logistics of data sharing. These include ensuring that health research datasets are accessible and usable for researchers and other users, and maximizing the potential to link datasets—those collected for the purpose of research with those collected for other purposes.

MAXIMIZING ACCESS AND RE-USE OF RESEARCH DATA: LESSONS ABOUT OPPORTUNITIES AND CHALLENGES FROM THE SOCIAL SCIENCES

Myron Gutmann, professor of history and director of the Institute for Behavioral Sciences at the University of Colorado Boulder, and former assistant director for social, behavioral, and economic sciences at the U.S. National Science Foundation, gave a keynote talk on lessons from the social sciences.

A Bit of History

According to Gutmann, sharing of data has been going on in the social sciences for decades. The Roper Center, now located at the University of Connecticut, was founded by the Roper Organization in 1946 to enable sharing of polling data. The Inter-university Consortium for Political Research, currently the Inter-university Consortium for Political and Social Research (ICPSR), was founded in 1962. In the early 1960s, census micro data and major surveys began to be shared and these publicly available data became the basis for research in the social and behavioral sciences. Several factors, including the rise of social science disciplines and nongovernmental research, coupled with an interest in social science immediately after World War II, led to an interest in increased understanding of social processes and investment in social science. The availability of new computer technology in the 1960s enabled researchers to actively use and share data about society.

At the same time, Gutmann explained, networks of repositories of data rapidly began to form. Initially, they were national in scope and very specialized, but soon, like ICPSR, broadened to an international scale. The Council of European Social Science Data Archives, almost as old as ICPSR, is an important European network. The creation of new mechanisms for technology, including the Web, led to virtual networks. The virtual data center at the Institute for Quantitative Social Sciences at Harvard University has spurred important activities in cross-archive networking. Access to data available through the ICPSR doubled with the advent of the Web. Gutmann said many social science surveys operate with the assumption that data will be publicly available within a short timeframe, which makes the surveys important resources.

Data Sharing: U.S. Context and Literature

At the same time, there has been growing scientific opinion and discourse tied to issues related to data sharing, Gutmann said. Studies at the U.S. National Academy of Sciences[1] have addressed issues such as privacy and confidentiality (see, for example, *Protecting Participants and Facilitating Social and Behavioral Sciences Research* [2003]). *Conducting Biosocial Surveys* (2010) discussed confidentiality issues in detail, while *Putting People on the Map* (2007) discussed the complexity of sharing data that have specific geographic locations in them. Standards for metadata have helped make it possible to understand and combine data but doing

[1]These reports and others from the National Academies of Sciences, Engineering, and Medicine, published by the National Academies Press, can be downloaded for free at http://www.nap.edu/.

so requires resolution of sometimes complex confidentiality and privacy issues.

One indication of the current acceptance of data sharing in the United States is a policy issued in 2013 by the U.S. Office of Science and Technology Policy, which called for all federal agencies that support research to make data and publications publicly available in a timely manner. Developments in the production and analysis of data will continue and will raise new issues. The availability of social media data, the increasing role of administrative data, and the use of commercial data in research create a broad innovation space for creative analysis of social problems, Gutmann said. They also lead to analytic issues tied to the size and complexity of integrated data and how to use them in a meaningful way, as well as how to know what inferences can be drawn from combined data that may have uneven coverage and uneven quality.

Confidentiality Protections

An overriding consideration is confidentiality protection. Gutmann suggested the answer to how to protect people is to have a graduated system that provides various means of protection, with the access difficulty increased as the risk of disclosure increases. For example, very simple data with little risk of disclosure could be accessed via the Web. For complex data with very little risk of disclosure but some risk of harm, people could sign a contract not to share these data. There may be cases where there are data with a very high risk of disclosure or a very high risk of harm. There, he suggested, the answer may be an enclosed data center. There are all sorts of ways of dividing up the gradient and providing a wall between a potential intruder and the data, he said, as well as other efforts such as to limit data, alter the data, provide secure access, and simulate data.

Gutmann said an approach commonly talked about in census data samples is swapping. For example, for two people in two locations who are quite similar (e.g., age, number of children, etc.) where only the income is different, the income is "swapped" to make it harder for an intruder to really know whom they had found.

Looking Ahead

Gutmann reported that in his own work looking at the relationship between agriculture, population, environment, social change, and health in the United States, he is seeing a move toward large integrated data collections. What that means, he said, is that single repositories are impossible to imagine as being the only solution.

In his view, this requires new kinds of infrastructure and strategies for confidentiality protection. Although one of the advantages of distributed data is that the linkable information is not all in the same place, it also becomes much harder to get to and the analytic requirements are large. "We have to think then about how we are going to deal with scale and how we are going to deal with human data reporting," he said. "Yet, if we want to bring them together at a global scale, we are going to need to find a way to do this." On top of that, he said, data sharing will require thinking about the policies, laws, and culture that vary nationally and sometimes regionally (e.g., the European Union) or locally.

Recipes for Long-Term Success

Gutmann suggested examples for long-term success exist in the social sciences. "It requires that we engage communities. It requires that we have high-quality data collection and management. It requires that we have rapid and easy data sharing. It requires that, most of all, we . . . build capacity steadily, rapidly by training data users, training data prep managers, training everyone involved, bringing together these engaged communities to talk about the important problems that we solve," he stated.

A participant posed the question of which stakeholders should play which roles. Gutmann responded he sees a distributed and divided responsibility and everyone has a role to play. Institutions need to define a policy sphere in which to get effective research done. Communities have the role "which is both to do our jobs and to do them well, but also to keep pushing our institutions to understand that we cannot solve the problems we need to solve without their taking an active role," he said.

In response to another participant's question, Gutmann commented that to the extent that a culture can be created to share data, it should happen. His position is that data should be made as public as possible. But, he pointed out, the incentives for research subjects and researchers have to be in place. Related to this, he observed most of the workshop participants have very specific interests in improving health in their communities, and that may provide a different set of incentives than those of, for instance, an assistant professor of political science in the United States or in Northern Europe.

ENABLING DATA LINKAGE TO MAXIMIZE THE VALUE OF PUBLIC HEALTH RESEARCH DATA: A PHRDF REPORT

Felix Ritchie, a professor of applied economics at the University of the West of England (UWE) in Bristol, and his co-author, Alex Montgomery,

a technical officer at DataFirst[2] at the University of Cape Town, presented a recent Wellcome Trust–published PHRDF report on enabling data linkage.[3]

Ritchie began by saying that the PHRDF report endorses almost everything raised in the plenary presentation. It draws primarily from high-income countries (HICs) because there is very little in the literature about low- and middle-income countries (LMICs). To try to address this gap, the report includes case studies based on interviews with people who work in LMICs. The aim of the report was to think about how data linkage could boost public health research and the barriers to useful data linkage. He said it focuses on what is practical and useful, rather than being exhaustive.

The project team represented the business school and the public health school at UWE, DataFirst, and the Center for Injury Prevention Research in Bangladesh. It was designed to offer a range of perspectives by including a mix of people from different socioeconomic backgrounds and work perspectives, including data access, clinical work, and epidemiological work.

The project included a nonsystematic literature review, formal and informal interviews (face-to-face, telephone, and via Skype), and the subset of interviews that served as case study examples. These methods were supplemented with the team's own experience and personal understanding of data access.

Key Findings of the PHRDF Report

Ritchie summarized the main findings of the report:

- *Change the tone of the debate*. According to Ritchie, "data should not be used for research or linked unless it can be done safely and securely" and "data should be available for research and linking unless it cannot be done safely and securely" functionally mean the same thing. But the default of the first statement is that data sharing is closed and the default of the second is that it is open.

[2]DataFirst was set up by the Mellon Foundation Fund to share survey and administrative micro data and increase skills among African researchers. It is an ongoing repository and curates data to international standards. DataFirst efforts include versioning, quality control, and disclosure control, and it provides an online help desk to people who use the data in their repository. The group also runs workshops to train data analysts. For additional information, see https://www.datafirst.uct.ac.za/ [August 2015].

[3]The full report can be found at http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTP056860.htm [August 2015].

He suggested that the debate needs to be shifted from a default-closed position to a default-open position.

- *Policy decisions need to be more evidence based*. Researchers using data for research purposes are generally viewed as low risk, but things can still go wrong. In the wider context of protection for data, academic research use of data is a low-risk activity, but that message is not widely understood outside the data community. The academic literature, which guides decision makers, focuses on intruders. However, usually when problems occur, it is because someone made a mistake.
- *Narrow informed consent is not enough for good epidemiological research*. When talking about linked data, a lot of data that might be linked has not been collected with consent for research. For example, administrative data or census data were not provided with consent for use in statistical research. Often, when data are collected for health research purposes, broad consent is acquired for that data. Ways are needed to link data where there is not consent or it is not sensible to acquire consent when used for epidemiological research.
- *Maintaining good relationships is the key.* The project's case studies showed that good relationships are the key, Ritchie said. If the stakeholders, ethics committees, and users are on board at the beginning of any data-sharing project, everything goes much more smoothly.
- *Incentives to manage and share data are weak*. Data funding is often tied with research funding, but the interest is in getting the research out. The incentives to work on data are weaker.
- *Different things happen in different places*. There may be a hierarchy of problems, from data to organization to institution, and the problems experienced differ for HICs and LMICs. In some places, the issue is getting access to data; in others, it may be linking the data, getting useful data, or being able to use the data held in a different system.

High-Income Country Experience

Ritchie reported that from the perspective of the HICs, there are problems with data and with organizational and operational issues, but institutional issues are dominant, such as relationships with the people who deposit the data, with ethics committees, and with the general public, and unrealistic risk assessment and worst-case scenario planning may result. According to Ritchie, stakeholder management works well in HICs. Stakeholder management involves both early planning and education/

communication. In HICs, it includes talking with ethics committees when they do not have expertise in the topic at hand.

Low- and Middle-Income Country Experience

There is less information about LMICs and data sharing, according to Ritchie, but based on the information they had, the LMIC experience is dominated by operational and quality issues. For example, publicly funded health data are held by institutions and universities, and only available to collaborators, rather than having research facilities in place to share those data. Although international funders have data-sharing requirements, there are not very strong data-sharing requirements for national funding bodies, and data-sharing requirements of international bodies are not enforced. LMIC experience tends to be focused on "pools of expertise" rather than a critical mass of researchers, to ensure robustness and enable sharing with colleagues.

Montgomery talked about the experience of several projects in South Africa, which is in a transition from a LMIC to a HIC, as a case study. South Africa has linkable data and a number of projects have been undertaken, including a project to link data from the Agincourt Health and Demographic Surveillance System with other data sources. Their data-linkage efforts ran into various operational and statistical barriers, and ethical concerns were raised. The Department of Health is trying to operationalize due diligence by setting up a preapproved database and procedures for using it to increase researcher access to the linked data.

Changing the Conceptual Framework

The Wellcome Trust's report *Enabling Data Linkage to Maximize the Value of Public Health Research Data* (2013) recommends changing the conceptual framework around data sharing. Ritchie observed commonality in the views of many participants and a lot of knowledge about what works, but the knowledge is sometimes in the wrong place and not accessible to decision makers. He suggested information needs to be made available to avoid continual reinvention.

A participant suggested the need for strong case studies where data linkage led to a finding that subsequently led to a policy change and impact. Ritchie responded with an example of a cohort study with a control group that was looking at the use of statins in the prevention of cardiopulmonary disease and stroke. It covered about a 4-year period and seemed to be showing a significant difference. The study was then extended by data linkage for a further 25 years. The discrepancy in mortality rates greatly expanded. In this case, an expensive cohort study was extended at virtually no cost to prove something quite remarkable.

Developing Practical Guidelines

The PHRDF report also recommended developing practical guidelines and measures. According to Ritchie, "everything has been solved; everything has been done somewhere." But, the available information has to be more widely known. Ritchie highlighted what he considered to be some primary issues for LMICs:

- Establishing research data infrastructure to support health data usage and linkages, such as DataFirst's free data position and archiving service for all African organizations.
- Building quantitative skills, expanding from having pools of experts to more critical mass.
- Data management, with data collection only part of the overall research process.
- Targeted funding for good data collection and curation.

Lynn Woolfrey, from DataFirst, stressed infrastructure is very important to support data linkage in African countries. The secure data service that enabled the project team to get data "would have never been in the research domain." Woolfrey pointed out the secure data service is open to all researchers, not just African researchers.

Discussion

Data Sharing/Linkage Case Studies

A participant observed that it may take a long time to address the concerns raised, but there are multiple examples of how data linkage can and has improved health or informed policy issues. He pointed to a Scottish study on a specific insulin medicine that put two datasets together, with immediate implications for practice. Another participant shared an example of linking population-based HIV sero conversion data obtained by household surveillance to antiretroviral treatment data from public clinics.

Ethical Review Boards

Participants had a lively exchange about the role and perspective of ethical review boards in reviewing and approving research involving shared data. Some felt that the review boards, also known as institutional review boards (IRBs), stand in the way of research involving data access or data linkage because of a lack of understanding, which can lead to an adversarial relationship. Others pointed out that ethics boards have their

own responsibilities and requirements; they are not being intentionally obstructionist. One of the report authors shared that some researchers in the study said they felt that if review boards feel the need to do a full review of an approved proposal, it can be taken as the board thinking the researcher is not competent. Another opined that ethics board reviews of secondary analyses should not have to go through a full ethics approval process if the data are available through archives developed for researchers.

A participant shared an example of research he had done with hypersensitive studies designed to re-identify people. His research team went to the IRB well before submitting an application for the research and conducted a series of workshops with the IRB members. He said it made a big difference and is the kind of thing needed when "doing anything sensitive."

Another participant echoed this sentiment, saying "you have to think about how you can help them to do their job in a way that they feel comfortable with." A participant representing an ethics review board reinforced this, suggesting an engagement strategy with the board and not just submitting a "complicated proposal and expect[ing] everybody to get up to speed." To provide an example, a participant shared that when HIV prevention trials were starting, the World Health Organization (WHO) ran global education programs for ethics committees to help inform them of issues and empower them to competently review these kinds of proposals. Another participant reported H3Africa does similar work. The Global Alliance for Genomics and Health[4] is also engaging in a process to think about ethics equivalency.

The discussion closed with a comment by a participant about the importance of changing mindsets of both ethics oversight bodies and funders. To engage in data linkage requires good access to data.

BUILDING PARTNERSHIPS FOR DATA SHARING, LINKAGE, AND RE-USE

This panel session picked up on earlier points about building partnerships with statistical authorities, data controllers, data owners, those responsible for the ethical control permission of research, research communities, and funders.

[4]For more information, see http://genomicsandhealth.org/ [August 2015].

The ALPHA Network

Basia Zaba, professor of medical demography at the London School of Hygiene and Tropical Medicine and head of the African Longitudinal Population-based HIV data on Africa (ALPHA) network, presented on the network.[5] ALPHA is a network of 10 community-based HIV study sites in Eastern and Southern Africa. All the studies existed before the network was formed and have their own independent scientific programs. Some have partners in the North; others are independent organizations. They came together because they realized that they were addressing similar questions in different ways.

Initially, most of the studies ran protocols that they call informed consent without disclosure. If subjects wanted to know their test results, they could obtain them, but the studies did not insist that people learn their HIV status as a result of participating in the testing programs. This has changed over time, and gradually some of the sites are moving toward an expectation that everybody who is tested will want to know her or his status in order to access treatment.

According to Zaba, the real strength of the ALPHA network is the follow-up with HIV-positive members. They have identified nearly 50,000 people who are HIV positive and have over 150,000 person years of follow-up with them. This enables researchers not only to do very in-depth analyses of mortality, but also to investigate risk factors for acquiring HIV and to do direct measurements of HIV incidents, which are not available from other sources.

Self-Interest as a Data-Sharing Imperative

The motivators for the ALPHA network's members to share data with each other "were the concerns of the people, our clients, who needed our data," Zaba said. The Joint United Nations Programme on HIV/AIDS (UNAIDS), WHO, and the Global Fund all wanted data from these sites, but they wanted to know that the data were generalizable. These sites were chosen because they are places of interest or places where it was feasible to do field work. If all the sites show similar results, then the people who need these results know that there is a far better chance of the results being replicable across a lot of settings.

Zaba commented that when they pool their data, the statistical power is much greater than if each site only has its own data to look at. But while all the sites may have looked at HIV and fertility "somebody has done it one way, somebody has done it another way," she said. By solving the

[5]For additional information, see http://alpha.lshtm.ac.uk/ [August 2015].

technical problems of data sharing together, "we are gradually learning also how to share our data with the big wide world outside." Trust is not really a problem for the network, she said.

Shared Data Resources

The network has built up an impressive shared data resource, Zaba said, which consists of the demographic episodes and events being experienced by the people followed; HIV test dates and results; and various linkages between the individuals observed (e.g., co-parenting, co-residence), sociodemographic data (e.g., education, marital status, sexual behavior), and data on verbal autopsies and cause of death. Africa has no death registration system or medical certification of cause of death. Data rely on the reports of people who have cared for the deceased during their final illness to describe the symptoms experienced, in order to get some idea of the likely cause. The network also has self-reported data from interviews in the community about HIV care and treatment and has recently embarked on linking with the clinics that provide treatment and care. The Masaka site has its own research clinic, but most ALPHA member sites use government clinics, which serve their populations and are gradually building up data linkages, he explained. The network has a current grant application into Welcome Trust to make their data tables publicly sharable.

Capacity Bootstrapping

Zaba described the ALPHA approach to capacity building, as "capacity bootstrapping." ALPHA has a scientific advisory committee that involves principal investigators from all the member study sites. It also involves people from the London School of Hygiene with statistical expertise and has other outside members from WHO, UNAIDS, and other organizations. The scientific advisory committee chooses research topics to do a literature review, looking at what different sites have done on that particular topic.

The members agree on a series of basic analysis and then very carefully specify a harmonized dataset that every site has to supply in order to do this analysis. The harmonized dataset is the minimum dataset that is required for the analysis. But it is also the lowest common denominator, in terms of the categories that all the sites can achieve. All the sites can recode their data to produce these harmonized datasets. They then organize workshops to discuss the public health rationale, the epidemiological theory, and the statistical theory for the analyses. Very importantly, the workshops are not just for the analysts. The ALPHA approach to capac-

ity building is also aimed at data managers, as well as more traditional epidemiologists and others who analyze the data.

They agree on what kind of joint analyses can be done and which might be using the pulled dataset, then they follow up with publications. ALPHA findings have been showcased in several special issues of journals. They are branching out into policy analysis and health facility studies with funding from the Gates Foundation and the Wellcome Trust.

Technical Issues: Harmonization, Documentation, and Data Linkage

Data harmonization has made their data more valuable, Zaba said. The process has helped them to understand what the users want. When the ALPHA projects started, the data collected were very simple, focused on incidence and prevalence trends. The data have gotten more complicated over time and now include clinical linkages. Harmonization gets more complicated as the data do.

In Zaba's view, "big data" constructs are more challenging in LMICs. For example, except in South Africa, there are very few indexing identity (ID) variables, like national IDs. There are no Social Security numbers or post office codes. While mobile phone use is common, there are no personal phones in rural Africa. There is not the same kind of personal identification of a phone number and an individual as in other locations. There is also much less certainty about dates of events and far more variability in the rendition of names, even name order.

According to Zaba, there is a lack of statistical theory for linkage failures. In addition to missing links, they also sometimes have real multiple links because somebody came back into one of the studies and was not identified as a returning migrant, causing additional statistical challenges.

Collaborations

The ALPHA network collaborates with other networks, including INDEPTH, Idea (a network of HIV clinical cohorts), and the HIV modeling consortium. A participant observed 8 of the 10 ALPHA sites are also in INDEPTH and asked about the overlap and motivation for having two networks rather than extending one of them. Zaba responded that INDEPTH is older than ALPHA, and some of the ALPHA sites joined INDEPTH after being in ALPHA and learning of the benefits of INDEPTH. ALPHA is very specialized in that all study sites do community-based HIV surveillance. They are hoping to use the INDEPTH data archive.

Zaba closed by emphasizing funding problems. The sites that contrib-

ute the data "are not the favorite for most funders. They are sort of boring workhorses of the research world, rather than open-ended, basic observational studies," she said, noting they are the platform for other research projects. They may have to tax the projects that build on their platforms in order to make sure that the health and demographic surveillance system studies survive. Zaba also commented on the lack of mobility of her employees—researchers from the North can easily work for years at a time in most African countries, but researchers from African countries find it difficult to get employment permits to work in each other's countries and almost impossible to work in the North.

The WorldWide Antimalarial Resistance Network (WWARN)

Karen Barnes, professor of clinical pharmacology at the University of Cape Town and director of the pharmacology modules for the WorldWide Antimalarial Resistance Network (WWARN),[6] presented on WWARN's efforts to bring the antimalarial research community together to make malaria treatment more effective.

WWARN was created with the mission of providing the information necessary to prevent or slow antimalarial drug resistance, to make sure individuals have the most effective treatments, and to thereby prevent malaria morbidity and mortality. The network met for the first time in 2004, got their first planning grant from the Gates Foundation in 2007, and became firmly established in 2009. She observed that much of their thinking happened in parallel or before issues around data sharing.

Their efforts focused on the malaria community in malaria-endemic countries that had seen drug resistance to chloroquine and then sulfadoxine-pyrimethamine. They had very clear data that it took between 4 and 8 years between having clear evidence of failing treatment to having a change in policy implemented. During that time, they estimated about 112,000 extra deaths happened each year as a result of staying with failing drug policies.

She pointed out WWARN was created to detect resistance early and speed up the time between when resistance is known to development of a more effective treatment policy. That sort of mandate, she said, made it easy to attract people with similar goals to work together. There are currently over 230 partner institutions across the globe, with leadership of WWARN across many sciences. Cape Town leads the pharmacology module and Bangkok leads quality assurance/quality control. Oxford is the hub, but the project is very global. Barnes reported that one achievement is having over 100,000 clinical trial patients in malaria studies. Two-thirds

[6]For additional information, see http://www.wwarn.org [August 2015].

of the artemisinin combination therapy data published to date are already in the WWARN data repository.

WWARN works with the malaria community to collect data on the clinical efficacy of drugs, the molecular markers associated with antimalarial drug resistance, as well as in vitro data of drug resistance. One arm looks at pharmacology to separate out poor-quality drugs from true resistance.

WWARN has an ability to link data in all those domains and to link metadata at all study sites. At first, it was considered a major deterrent if WWARN was specific about how the data needed to be submitted—the approach was to take what they could and make it their problem to make it compatible. Oxford helped ensure that the bioinformatics technology was secure. The project spent a lot of time on curating data and checking data quality, which automatically gave feedback to the people contributing their data. In return, the contributors got very detailed reports generated often in a matter of weeks, which they could then use to publish their data much more quickly with tables, graphs, and other depictions.

Data are standardized so that they can be reanalyzed. Publications are not the primary goal. Barnes pointed out that WWARN provides numerous free tools to help researchers with planning their data; provides templates, protocols, and tools for analyzing data; and generates automated reports.

WWARN also runs a proficiency-testing program for laboratories, pharmacology and in vitro drug quality laboratories. They send out samples and ask what concentration recipients think is in the sample; the results are anonymized. She reported that with each round, people are getting closer and closer to the target concentration. She also noted that the laboratories in the North did not outperform the laboratories in the South.

Data Visualization

WWARN works on presenting data visually, such as maps, to allow policy makers, health care workers, and others see what the data mean so they can see the need to change policies, which is the ultimate goal. Data visualization makes the findings more accessible and in interactive format, allowing the data to be interrogated in specific locations. A map can illustrate, for example, that ASP resistance is a problem in East and Southern Africa, but less of a problem in West Africa. Data on drug quality can also be visualized, with an indication on a map of every place that had a report of a substandard counterfeit antimalarial.

A Success Story

WWARN's hardest challenge is to slow resistance and develop the antimalarial drug pipeline. The next wonder drug is at least 5 years away, she said, so the task is to make the current drugs last for as long as possible. They have worked to identify factors to promote resistance, optimize regimens, and then target interventions appropriately.

For example, the dosing of dihydroartemisinin-piperaquine in young children shows the promise of their work and the value of their approach. This drug is the artemisinin combination that is currently available that has the best potential to last until new drugs are available. They pulled the available efficacy study data done with this drug, and the main result showed almost a 98 percent cure rate.

Because they had such a large dataset linked to enough other details, they were able to identify that the youngest children, ages 1 to 4, who in endemic countries are those without immunity, had a four-fold higher risk of recrudescence. They also knew that the drug concentrations were generally lower in this population at the currently recommended dose. They were able to determine that increasing the recommended dose slightly could halve the risk of treatment failure, and achieve the WHO goal of more than 95 percent cure rates.

Barnes acknowledged concerns about the potential risks of "just pushing up a dose." To address that, they pulled all the pharmacokinetic data of drug concentrations that had been measured, from all WWARN sites that had measured them, and modeled how to shift the dose to make sure that the minimum exposure was high enough, but the maximum exposure was not too high. They provided their modeling to WHO, which has changed the treatment guidelines that will come out this year, to include a higher recommended dose for children ages 1 to 4. The project hopes that the increase will enable the field to hang on to this useful drug longer.

Prioritizing Resources

The availability of large amounts of data in a variety of areas helps WWARN to identify where more data are needed and enables them to use data to support effective interventions. For example, sulfadoxine-pyrimethamine is not generally recommended as a treatment but is used as preventive treatment in young children in areas with seasonal malaria transmission and in pregnant women. By mapping resistance rates, WWARN can help target studies in the right places, increasing efficiency. The project has the potential to preserve the efficacy of available antimalarials working on optimizing dosing regimens, Barnes said. They are particularly worried about dosing for malnourished children at the moment.

Barnes concluded by saying that as a data center, WWARN has developed the scientific and ethical rationale for data sharing and has the potential to provide long-term secure data storage and to help their data contributors meet the requirements of journals and of regulatory agencies. Their work can help make drugs last longer and be used to help inform the development of new drugs. The assumption in the past has been that the same milligram-per-kilogram dose is going to work for everyone, which is not the case. WWARN provides accurate, useful intelligence to inform this process.

Barnes observed that WWARN started at a time when people were not expected to share data. In WWARN's participatory mode, people are less concerned that their data will be misused in secondary analysis, but WWARN does provide expertise and insight to make sure that secondary analyses are valid and reliable. She reflected that there is capacity building at every level and that the feedback on data, training programs, and tools lifts up everyone's quality of work, North and South. Taking initiatives like WWARN forward will depend on giving people support to avoid further separation of established and emerging researchers, she said.

Statistics South Africa

Dan Kibuuka, a director at Statistics South Africa (Stats SA)[7] responsible for health statistics and for managing the acquisition, collection, and analysis of health information from household-based surveys, talked about what Stats SA does and what they are capable of doing, and the interactions that have taken place. He reflected that the issues discussed at the workshop are different in the government context. His data-collection work is supported by law. Stats SA is a government department tasked with collecting, processing, and analyzing data. They assist other government departments that do not have the capacity to collect data through household surveys since they have an infrastructure that reaches to the remotest district of the country. They also assist nongovernmental organizations and other government departments in the collation and analysis of administrative data.

Before a survey is conducted, Stats SA holds user consultation workshops. They consult with the National Department of Health in particular to determine the current issues in the country before going to the field with questions. Stats SA has several surveys with health data:

- General House Survey (annual)
- Living Conditions Survey (every 5 years)

[7]For addition information, see http://www.statssa.gov.za/ [August 2015].

- Income and Expenditure Survey (every 5 years)
- Community Survey (every 5 years)
- Census (supposed to be every 5 years, but funding limitations have sometimes prevented that and required a large community survey instead)

Stats SA works with the National Department of Health, which has administrative health data, especially data from the District Health Information System (DHIS), to compare data they collected through surveys with data from the DHIS. The Department of Home Affairs provides Stats SA with death certification information that Stats SA analyzes and uses to produce an annual report. Stats SA is also coordinating with colleagues from the South African Medical Research Council and the Human Sciences Research Council to pool resources. They will together run the South Africa demographic survey.

Stats SA collects data on such topics as disability, immunization coverage (beginning in 2016), childhood diseases, causes of death, and out-of-pocket medical expenses, with sample sizes of about 93,000 people. They have been able to respond to stakeholder issues by adding questions to existing surveys. For example, data on immunization coverage was added for the Department of Health because of concerns raised by WHO. Similarly, information on childhood diseases was added for the Department of Health, with guidance from them on how to ask questions.

The combined data are a rich source of information, he said. He specified a few factors that enable sharing, linking, and re-use:

- Know which health data are being produced by whom, then find out who collects the data and explain how you want to use the data.
- Know the frequency of production. Researchers can plan publications or other targets based on the data-production timelines.
- Know the procedures for acquiring and accessing the health data.
- Know what can be offered through networking and collaboration.

Data linkage at Stats SA is in its infancy, but they are evolving to a more sophisticated linkage model that involves working to move beyond basic descriptive reports and aiming to do more analyses. They are exploring whether they can link their health survey data to causes of death, for example. He welcomed researchers to offer partnerships with Stats SA and to offer research skills in exchange for data access. Stats SA also has a well-structured data repository to support data re-use.

Kibuuka shared his view that three things can make partnerships fail: competition, lack of trust, and funding.

Discussion

Fundraising/Core Costs

A participant commented that the examples presented illustrate the power of pooling data and asked why fundraising is difficult. Zaba responded that repeated demographic surveillance is not viewed as cutting-edge research, and funders seem primarily interested in intervention trials. Zaba also commented many sites are used to being independent and "don't want to limit themselves just to being funded for a narrow common range of questions that are only answerable through the network." Barnes responded that there is more opportunity for start-up funding, but most projects do not have the core funding that makes that possible. Researchers use the data largely because they are available for free, yet society and public health really benefit from data sharing, she said. A participant commented that the significant operational costs of the core elements should decrease over time as efficiencies are gained, but that is not what he sees with funding requests. There may be a tension between maintenance of an established operation and continuing to grow, he posited.

Sustainability/Research Culture

Training the next generation of researchers to sustain research such as linkage to use of administrative data is a challenge, observed a participant. It is more attractive for young researchers to develop their own cohort studies, collect their own data, and publish an original paper, and asked how to change that mindset. Zaba commented that ALPHA's capacity building and training addresses what academic training cannot because it extends beyond theory and focuses on how to approach an analytical solution to a specific problem. She expressed optimism that the people trained will be able to continue the type of analysis they were taught, if funding is available.

Barnes connected the solution to the benefits possible from sharing data that are not possible through individual studies. "I think that excitement of what more you can do without having to start studies from scratch . . . will sustain the new paradigm shift," she observed.

6

Next Steps: Maximizing the Use of Data to Improve Public Health

Steven Kern of the Bill & Melinda Gates Foundation moderated the final session. He summarized the highlights of the workshop (see Chapter 1) and, by reframing a slide presented by Kobus Herbst at the start of the workshop (Figure 2-1), illustrated two key perspectives that he commented permeated the discussion:

1. Data users and data collectors should work in tandem with one another, providing stability to multiple components of the data-sharing continuum (see Figure 6-1) and
2. Consideration of issues related to the data collected, such as confidentiality, avoiding harm, and potential benefits to data subjects, needs to be paramount (see Figure 6-2).

Ideas for Next Steps

In the course of the workshop, several presenters outlined their ideas for next steps needed to facilitate data sharing in Africa, and breakout groups throughout the workshop also shared specific ideas. Many participants expressed support for the benefits from sharing data, but also said that work is needed to ensure that data sharing is enhanced. Kern summarized what he understood to be possible next steps:

- *Relationships*. The importance of relationships to ensure trust and confidence in the research was raised by many presenters and

participants throughout the workshop—relationships with participants, researchers in other institutions, ethical review boards, and the communities in which research takes place. Clarity about the purpose and intended benefit of research is critical to each of these relationships. Community engagement strategies to develop trust and confidence of community members and training for ethical review boards can help build needed relationships.

- *Tone.* "Researchers and research institutions need to change the tone on data sharing to make it the default assumption of research," Kern said. Researchers should think of themselves as stewards and custodians of data, rather than owners of data. Changing the tone will require resources and case studies that illustrate the benefits of sharing.
- *Capacity and infrastructure.* Sharing data requires the availability of quality data. Several participants raised the need for training in data management, curation, and analysis to ensure that quality data are available to be shared. Partnership and collaboration agreements should have capacity built in, either through explicit arrangements (fair contracting) or through "bootstrapping." African institutions face a major challenge in finding resources to build capacity given the limits on core functions available in typical funding arrangements.

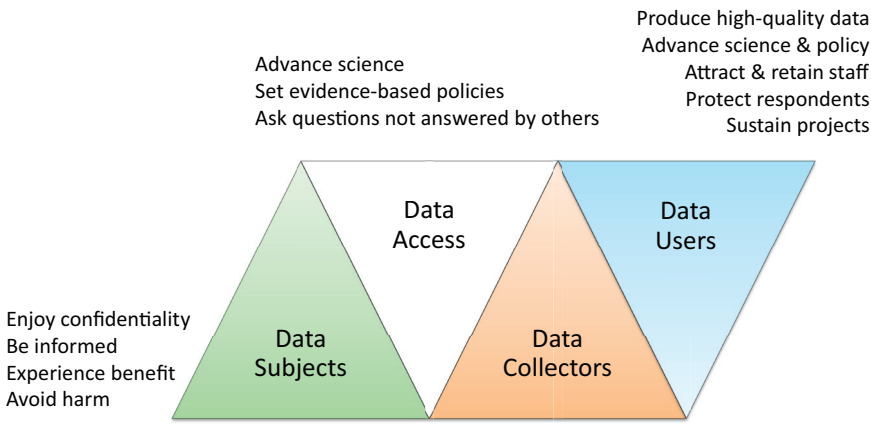

FIGURE 6-1 Data continuum players as equals.
SOURCE: Adapted from Herbst, K. (2002). Wider accessibility to longitudinal datasets: A framework for discussion. In National Research Council, *Leveraging Longitudinal Data in Developing Countries: Report of a Workshop* (p. 43). Workshop on Leveraging Longitudinal Data in Developing Countries Committee, Committee on Population. V. L. Durrant and J. Menken, Eds. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

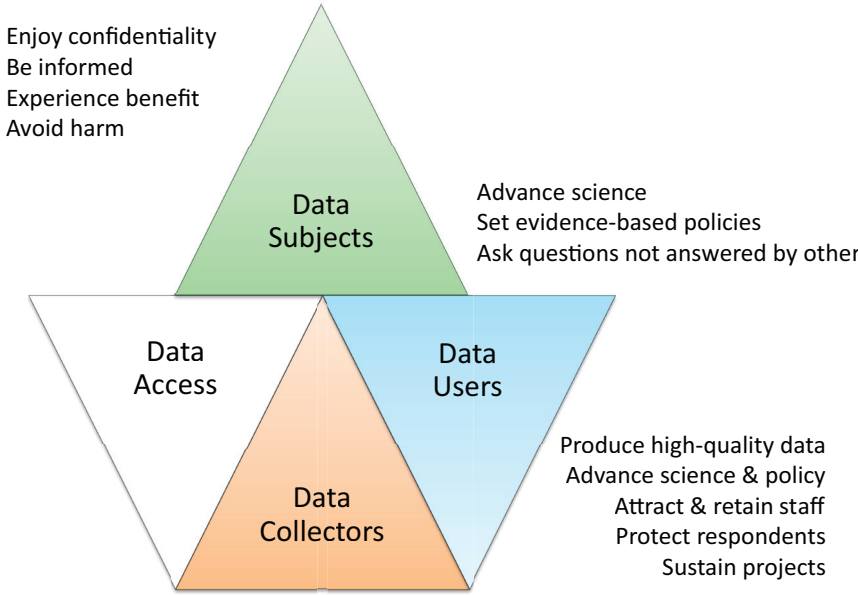

FIGURE 6-2 Data subjects as preeminent concern.
SOURCE: Adapted from Herbst, K. (2002). Wider accessibility to longitudinal datasets: A framework for discussion. In National Research Council, *Leveraging Longitudinal Data in Developing Countries: Report of a Workshop* (p. 43). Workshop on Leveraging Longitudinal Data in Developing Countries Committee, Committee on Population. V. L. Durrant and J. Menken, Eds. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

- *Incentives and recognition for data sharing.* The existing system of rewards and recognition for academics does not encourage data sharing. While the potential to improve public health is an incentive, more tangible incentives are needed. Metrics based on data publications, such as those available through Ubiquity Press, could be developed. Researchers should in general be acknowledged and rewarded for producing datasets. The value of research based on shared data needs to be acknowledged to attract young scientists.
- *Safety of data and ethical standards for sharing.* The need for appropriate confidentiality protections and ethical sharing of data is key. The approach to ensuring confidentiality can be tailored to the type of data and risk of exposure. Models for protecting confidentiality from the social sciences can provide a useful framework.

- *Resources and guidelines.* Information and examples provided at the workshop illustrated that much is already known about how to address data-sharing issues. At the same time, many participants said that information needs to be more widely shared, and several called for new resources (e.g., centers of excellence) and guidelines or models of good practice for funders and ethics boards, for the ethical use of data.
- *A broadened conversation.* The conversation needs to be broadened in a sustained way beyond those at the workshop, several presenters suggested. New relationships with groups like the World Medical Association and the West African Organizations for Health might help advance solutions. In support of this issue, David Carr suggested that participants use the workshop summary "as a basis for taking forward discussions with your own communities and stakeholders."

Future Challenges

Kern concluded with the observation that the future of public health research rests in part on the ability to maximize the use of data through sharing and linking data. Developments such as the explosion of online data via cell phones and the Internet, the emergence of citizen science, and the availability of increasingly complex data such as genomics indicate how the world is changing and present an opportunity to develop new and creative ways for research to respond. Finding solutions to enable regular and effective data sharing in Africa is an opportunity for scientists to be proactive and act at the cutting edge of ethics and science.

Appendix A

Workshop Agenda

SHARING RESEARCH DATA TO IMPROVE PUBLIC HEALTH

SUNDAY, MARCH 29, 2015

Session 1: **Introductory Session**
Chair: David Carr, Wellcome Trust

3:30 p.m. **Registration and Coffee**

4:15 p.m. **Welcome from Sponsors and Hosts**

4:30 p.m. **Meeting Context and Objectives**
David Carr and Katherine Littler, Wellcome Trust

5:15 p.m. **Keynote Presentation:** Benefits and Challenges of Data Sharing in the African Context
Kobus Herbst, Africa Centre for Health and Population Studies

6:00 p.m. **Plenary Discussion:** Priority Issues and Challenges

MONDAY, MARCH 30, 2015

Session 2: Establishing Equitable Terms for Data Sharing
Chair: Robert Terry, WHO-TDR

- How can we ensure that the researchers who collect and share data receive due recognition and reward for their efforts?
- How can we build capacity and ways of working to enable datasets to be used to their fullest potential to address health care and societal challenges across Africa?
- How should international collaborations and partnerships be structured to ensure equitable sharing of benefits?

9:00 a.m. **Session Keynote:** Establishing Equitable Terms for Data Sharing
Steve Tollman, University of the Witwatersrand, Johannesburg

9:30 a.m. **Panel Discussion:** Presentations (15 min.) and Moderated Discussion (30 min.)
Catherine Kyobutungi, African Population and Health Research Center
Michele Ramsay, University of the Witwatersrand
Jacintha Toohey, COHRED
Caroline Wilkinson, Ubiquity Press

11:00 a.m. **Break**

11:15 a.m. **Breakout Group Discussion:** Defining Fair Access Terms and Establishing Incentives

12:15 p.m. **Plenary Feedback**

12:30 p.m. **Lunch**

Session 3: Exploring the "Ethical Imperative" for Data Sharing
Chair: Katherine Littler, Wellcome Trust

- How can we balance the case for sharing data provided by research participants, with the need to safeguard privacy and confidentiality?

- How much do we know about the expectations of research participants, ethics committees and other key stakeholders?
- How can we build effective governance models to enable data sharing while managing risks and instilling trust?

1:30 p.m. **Session Introductory Keynote**: The Ethical Imperative for Data Sharing
Mike Parker, University of Oxford

2:00 p.m. **Stakeholder Perspectives on Data Sharing in Low and Middle Income Countries: Findings of a Multi-Site Study**

- **Ethical Challenges and Views About Best Practices in Data Sharing: The International Context**
  Mike Parker and Susan Bull, University of Oxford
- **South African Stakeholders' Perspectives About Ethics and Best Practices in Data Sharing**
  Doug Wassenaar and Colleagues, University of KwaZulu Natal
- **Kenyan Stakeholders' Perspectives of Ethical Issues and Best Practices in Data Sharing**
  Vicki Marsh and colleagues, KEMRI Wellcome Trust Research Programme
- **Implications of African and Asian Stakeholders' Perspectives for Ethical Best Practices in Data Sharing**
  Susan Bull, University of Oxford

3:15 p.m. **Questions and Discussion**

3:30 p.m. **Break**

3:45 p.m. **Breakout Group Discussions of Implications and Next Steps**

5:00 p.m. **Feedback and Plenary Discussion:** Ethical Best Practices in Data Sharing: Recommendations and Next Steps

5:30 p.m. **Reflections on the Day's Discussions**

6:00 p.m. **Close of Session**

TUESDAY, MARCH 31, 2015

Session 4: Enabling Data Discoverability, Linkage and Re-Use

- How can we ensure health research datasets are accessible and useable for researchers and other users?
- How can we maximize the potential to link datasets (including those collected for the purposes of research and those collected for other purposes) safely and securely—at a local, regional, and international level?

9:00 a.m. **Session Keynote: Maximizing Access and Re-Use of Research Data: Lessons About Opportunities and Challenges from the Social Sciences**
Myron Gutmann, University of Colorado Boulder

9:30 a.m. **Enabling Data Linkage to Maximize the Value of Public Health Research Data: Launching a New Forum Report**
Felix Ritchie, University of the West of England
Lynn Woolfrey, DataFirst, University of Cape Town

10:30 a.m. **Break**

10:45 a.m. **Panel Discussion: Building Partnerships for Data Sharing, Linkage and Re-Use**
Basia Zaba, London School of Hygiene and Tropical Medicine
Karen Barnes, University of Cape Town
Dan Kibuuka, Statistics South Africa

Session 5: Closing Session: Maximizing the Use of Data to Improve Public Health—Ways Forward

12:00 p.m. **Moderated Plenary Discussion:** Summary of Key Points from Across the Two Days and Key Priorities
Chair: Steven Kern, Bill & Melinda Gates Foundation

1:00 p.m. **Lunch**

2:00 p.m. **Close of Meeting**

Appendix B

Participants

First Name	Last Name	Company	Country
Dissou	Affolabi	National Hospital for Tuberculosis and Pulmonary Diseases	Benin
Mary	Ari	Centers for Disease Control and Prevention	USA
Karen	Barnes	University of Cape Town	South Africa
Bassirou	Bonfoh	Centre Suisse de Recherches Scientifiques Côte d'Ivoire	Côte d'Ivoire
Sarah	Bowles	Wellcome Trust	UK
Susan	Bull	Ethox Centre	New Zealand
David	Carr	Wellcome Trust	UK
Jantina	De Vries	Department of Medicine, University of Cape Town	South Africa

First Name	Last Name	Company	Country
Spencer	Denny	University of KwaZulu Natal	South Africa
Ali	Dhansay	South African Medical Research Council	South Africa
Roseanne	Diab	Academy of Science of South Africa	South Africa
Bai Lamin	Dondeh	Medical Research Council Unit, The Gambia	Gambia
Kenneth	Ekoru	International Health Research Group-University of Cambridge	UK
Peter	Elias	University of Warwick	UK
Timothy	Errington	Center for Open Science	United States
Melvyn	Freeman	National Department of Health	South Africa
Myron	Gutmann	University of Colorado	United States
John	Gyapong	University of Ghana	Ghana
Kobus	Herbst	The Africa Centre for Health and Population Studies, UKZN	South Africa
Irene	Jao	KEMRI	Kenya
Rachel	Jewkes	Medical Research Council of South Africa	South Africa
Sanjay	Juvekar	KEM Hospital Research Centre	India
Elizabeth	Kalama	KEMRI Wellcome Trust Research Programme	Kenya

First Name	Last Name	Company	Country
Mohamed	Karama	Kenya Medical Research Institute	Kenya
Steve	Kern	Bill & Melinda Gates Foundation	United States
Dan	Kibuuka	Statistics South Africa	South Africa
Jean-Francois	Kobiane	Institut Supérieur des Sciences de la Population	Burkina Faso
Brama	Koné	University Peleforo Gon Coulibaly of Korhogo and Centre Suisse de Recherches Scientifiques en Côte d'Ivoire	Côte d'Ivoire
Catherine	Kyobutungi	African Population and Health Research Center	Kenya
Katherine	Littler	Wellcome Trust	UK
Glaudina	Loots	Department of Science and Technology	South Africa
Stephen	Lwanga		Uganda
Dermot	Maher	Tropical Diseases Research/ World Health Organization	Switzerland
Vicki	Marsh	KEMRI Wellcome Trust Research Programme	Kenya
Honorati	Masanja	Ifakara Health Institute	Tanzania
Felix	Masiye	University of Zambia	Zambia
Adamson	Muula	University of Malawi	Malawi

First Name	Last Name	Company	Country
Germano	Mwabu	University of Nairobi	Kenya
Akindeh	Nji	Biotechnology Center, University of Yaounde I	Cameroon
Rebecca	Nsubuga	MRC/UVRI Uganda Research Unit on AIDS	Uganda
Thomas	Nyirenda	EDCTP	South Africa
Pierre	Ongolo-Zogo	Université de Yaoundé	Cameroon
Tolu	Oni		South Africa
Obinna	Onwujekwe	University of Nigeria Nsukka	Nigeria
Mike	Parker	University of Oxford	UK
Georgeanne E.	Patmios	National Institutes of Health/National Institute on Aging	United States
Colin	Pillai	Novartis Pharma	Switzerland
Bernadette	Ramirez	World Health Organization	Switzerland
Michèle	Ramsay	University of the Witwatersrand	South Africa
Felix	Ritchie	University of the West of England, Bristol	UK
Osman	Sankoh	INDEPTH Network, Accra	Ghana
Nelson	Sewankambo	Makerere University	Uganda
Blessing	Silaigwana	University of KwaZulu Natal	South Africa
Abdramane Bassiahi	Soura	ISSP, University of Ouagadougou	Burkina Faso
Robert	Terry	TDR-World Health Organization	Switzerland

First Name	Last Name	Company	Country
Steve	Tollman	School of Public Health, University of the Witwatersrand, Johannesburg	South Africa
Jacintha	Toohey	Council on Health Research for Development	South Africa
John	Vulule	Kenya Medical Research Institute	Kenya
Doug	Wassenaar	SARETI	South Africa
Richard	Wilder	Bill & Melinda Gates Foundation	United States
Lynn	Woolfrey	DataFirst, University of Cape Town	South Africa
Basia	Zaba	London School of Hygiene and Tropical Medicine	UK