National Cancer Institute

Think Tank on Identifiability of Biospecimens and -Omic Data

Bethesda, Maryland
June 11–12, 2012

SUMMARY

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

This Think Tank summary was prepared by Frances McFarland Horne, Silvia Paddock, Chandra Keller-Allen, Melanie Lymon-Harris, and Rose Li, Rose Li and Associates, Inc. under subcontract to SAIC-Frederick, Inc. This work has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The statements, conclusions, and recommendations contained in this document reflect opinions of the Think Tank participants and are not intended to represent the official position of the National Cancer Institute, the National Institutes of Health, or the U.S. Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

The authors thank Jane Bambauer, Laura Buccini, Michele Cargill, Deborah Collyar, Bob Gellman, Mark Gerstein, Tiffany Green, Pritty Joshi, Chris Kinsinger, Nicole Lockhart, Subha Madhavan, Leah Mechanic, Stefanie Nelson, Pearl O'Rourke, Laura Rodriguez, David Tabb, Jennifer Wagner, and Carol Weil for their review of earlier versions of this Workshop summary.

The Workshop planners gratefully acknowledge the contributions of all Think Tank participants, listed in Appendix 1, for their contributions to Think Tank deliberations and outcomes.

Suggested citation: Think Tank on Identifiability of Biospecimens and -Omic Data. National Cancer Institute, Bethesda, Maryland, June 11-12, 2012.
http://epi.grants.cancer.gov/workshops/identifiability/#summary

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

ANPRM	Advance Notice of Proposed Rule Making
AOL	America Online
CFR	Code of Federal Regulations
CLIA	Clinical Laboratory Improvement Amendments
CMS	Centers for Medicare and Medicaid Services
DAC	data access committee
dbGaP	Database of Genotypes and Phenotypes
dbSNP	Database of Single Nucleotide Polymorphisms
DHHS	United States Department of Health and Human Services
DNA	deoxyribonucleic acid
EHR	electronic health record
ELSI	ethical, legal, and social implications of medical research (Also referred to as ethical, legal and social issues in research)
GINA	Genetic Information Nondiscrimination Act
GWAS	genome-wide association studies
HIPAA	Health Insurance Portability and Accountability Act
ICs	NIH Institutions and Centers
IRB	institutional review board
IT	informatics or information technology
mRNA	messenger ribonucleic acid
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NHGRI	National Human Genome Research Institute
NIH	National Institutes of Health
PGP	Personal Genome Project
PHI	protected health information
PI	principal investigator
PLC	Portable Legal Consent

RNA	ribonucleic acid
SNP	single nucleotide polymorphism
TCGA	The Cancer Genome Atlas
WES	whole exome sequencing
WGS	whole genome sequencing

GLOSSARY[1]

Term	Definition
Advanced Notice of Proposed Rulemaking (ANPRM)	On July 26, 2011, the U.S. Department of Health and Human Services (DHHS) issued an Advanced Notice of Proposed Rulemaking (ANPRM) in the *Federal Register*, requesting public comment on how the current DHHS regulations protecting humans involved in research could be revised to improve efficiency and enhance protections for research participants.[2]
Anonymized	If all identifying information is irreversibly removed from biological samples and/or data following collection, the samples and data are said to be anonymized because it is no longer possible to trace or link them back to the individuals to whom they pertain. Advances in bioinformatics, genomics, and other high-throughput technologies raise the question of whether the anonymization of samples and data can be sustained in the long term.
Biospecimen	Biospecimens are samples of biological material such as urine, blood, tissue, cells, DNA, RNA, and protein, from humans, animals, or plants.
Coded	Biospecimens and data are considered "coded" when they are stripped of directly identifying information (such as name or Social Security number) and given an indirect identifier (i.e., a code) consisting of numbers, letters, symbols, or a combination thereof. Anyone who knows the key to decipher the code may be able to link individually identifying information to the biospecimens or data.[3]
Common Rule	Also known as the Federal Policy for Protection of Human Research Subjects, or Subpart A of the DHHS Regulations codified at 45 CFR part 46, the Common Rule was published in 1991 and includes the basic requirements for institutional review board review of research, minimization of research risk, and informed consent. It has currently been adopted by 17 Federal departments and agencies.[4]
De-Identification	The process by which data is stripped of information that would allow the identification of the human source of the data. De-identification may include preserving identifying information in a separate place, so that future linkage by trusted parties is possible.

[1] This glossary has been assembled to provide practical guidance on how workshop participants understood and used certain terms. The definitions provided herein are not necessarily comprehensive.

[2] The ANPRM can be accessed at http://www.gpo.gov/fdsys/pkg/FR-2011-07-26/pdf/2011-18792.pdf.

[3] See also: Guidance on Research Involving Coded Private Information or Biological Specimens (Office for Human Research Protections [OHRP]) at http://www.hhs.gov/ohrp/policy/cdebiol.html.

[4] http://www.hhs.gov/ohrp/humansubjects/commonrule/index.html.

Term	Definition
Epigenomics	Derived from the Greek, epigenome means "above" the genome. The epigenome consists of chemical compounds that modify, or mark, the genome in a way that tells it what to do, where to do it, and when to do it. The marks, which are not part of the DNA itself, can be passed on from cell to cell as cells divide, and from one generation to the next.[5]
Genetics	Genetics is a term that refers to the study of genes and their roles in inheritance, and explores how specific traits or conditions are biologically passed down from one generation to another. Genes (units of heredity) carry the instructions for making proteins, which direct the activities of cells and the functions of the body. Examples of genetic or inherited medical conditions include cystic fibrosis, Huntington's disease, and phenylketonuria (PKU).[6]
Genomics	Genomics, a more recent term than genetics, describes the study of all of a person's genes (the genome), including interactions of those genes with each other and with the person's environment. Genomics includes the scientific study of complex diseases such as heart disease, asthma, diabetes, and cancer, because these diseases are typically caused more by a combination of genetic and environmental factors than by individual genes. Genomics is offering new possibilities for more targeted therapies and treatments for complex diseases, as well as new diagnostic methods.[7]
Health Insurance Portability and Accountability Act (HIPAA)	HIPAA is a federal law affecting many aspects of health care access, coverage, and insurance.[8] HIPAA includes a Privacy Rule containing standards for minimizing the use and disclosure of specified categories of health information about individuals.
Identifiability	Identifiability is the potential ability to associate data or samples with actual persons. Identifiability references a spectrum of status points, from directly identifiable, to deductively identifiable, to virtually non-identifiable.[9]

[5] According to the Fact Sheet on Epigenomics published by the National Human Genome Research Institute: http://www.genome.gov/27532724

[6] Frequently Asked Questions About Genetic and Genomic Science. NHGRI. Available at http://www.genome.gov/19016904; accessed on November 7, 2012.

[7] See footnote 6 above.

[8] Detailed explanations of HIPAA can be found at http://www.hhs.gov/ocr/privacy/hipaa/understanding/index.html.

[9] Privacy, Confidentiality, and Identifiability in Genomic Research. Discussion document for workshop convened by the National Human Genome Research Institute, Bethesda, October 3-4, 2006.

Term	Definition
Institutional Review Board (IRB)	An IRB is a specially constituted independent review body comprised of medical, scientific and non-scientific members established and designated by an entity to ensure the protection of human beings recruited to participate in biomedical or behavioral research.[10] The Common Rule, codified in DHHS regulations at 45 CFR part 46, establishes membership criteria and rules for operations and review of research at 45 CFR 46.107 - 46.109.
Metabolomics	Metabolomics is the study of the biological metabolic profile of a cellular specimen in a specific environment at an isolated timepoint. This discipline depicts the physiological states of cells and organisms by focusing on carbohydrates, lipids, and other metabolites. Several analytical techniques are utilized to quantify the metabolic content of specimens such as mass spectrometry and electrophoretic applications.[11]
"-Omics"	"-omics" has been defined as "the study of related sets of biological molecules in a comprehensive fashion."[12] Examples of "-omics" disciplines include genomics, transcriptomics, proteomics, metabolomics, and epigenomics.
Protected Health Information (PHI)	The Health Insurance Portability and Accountability Act (HIPAA)[13] contains a Privacy Rule that addresses the use or disclosure of a subset of individually identifiable health information called *protected health information*, or PHI.[14] PHI includes identifiable demographic or other information relating to an individual's past, present, or future physical or mental health, or the provision or payment of health care to an individual, created or received by health care providers, health plans, employers, or health care clearinghouses.
Privacy	Privacy is a term used broadly in law, ethics, and health care to mean a range of concepts pertaining to freedom from unwarranted intrusion, decisional choice in personal matters, and preservation of confidentiality. In this summary document, the term "privacy" refers to an existing state in which access to information about an individual by unauthorized persons is prevented, or at least limited, by virtue of regulatory or institutional policy or contractual arrangement.
Proteomics	The study of the structure and function of proteins, including the way they work and interact with each other inside cells.

[10] Derived from the NCI Thesaurus found at http://ncit.nci.nih.gov/ncitbrowser/pages/home.jsf?version=12.09d.
[11] Derived from the NCI Thesaurus, see footnote 10 above.
[12] IOM (Institute of Medicine). 2012. *Evolution of Translational Omics: Lessons Learned and the Path Forward.* Washington, DC: The National Academies Press.
[13] http://privacyruleandresearch.nih.gov/pr_02.asp.
[14] http://privacyruleandresearch.nih.gov/pr_07.asp.

Term	Definition
Transcriptome	A transcriptome is the full range of messenger RNA, or mRNA, molecules expressed by an organism. The term "transcriptome" can also be used to describe the array of mRNA transcripts produced in a particular cell or tissue type.[15]
Whole Genome Sequencing	A procedure that can determine the DNA sequence for nearly the entire genome of an individual.[16]
Whole Exome Sequencing	The human genome consists of 3 billion nucleotides or "letters" of DNA. But only a small percentage—1.5 percent—of those letters are actually translated into proteins, the functional players in the body. The "exome" consists of all the genome's exons, which are the coding portions of genes. The term exon was derived from "EXpressed regiON," because these are the regions that get translated, or expressed as proteins, as opposed to the intron, or "INTRagenic regiON" which is not represented in the final protein. […] Exome sequencing offers a look into the genome that large-scale studies of common variation, such as the genome-wide association study (GWAS), cannot provide. GWAS can only identify variation in DNA that is common in the population, in at least 1 percent of people. But sequencing determines every letter in a DNA sequence, not just the ones known to vary, so it can reveal rare mutations that GWAS wouldn't uncover. Exome sequencing is a good choice for scientists today who are looking for rare mutations, especially when used as a complement to studies of common variation like GWAS. But as whole-genome sequencing becomes cheaper, that technique will likely be employed instead because it offers a look at all portions of the genome, not just those that include instructions for making proteins.[17]

[15] According to *Nature*'s "Scitable" at:http://www.nature.com/scitable/definition/transcriptome-296.
[16] From the NCI Thesaurus, see footnote 10 above.
[17] What is exome sequencing? Blog hosted by the Broad Institute at http://www.broadinstitute.org/blog/what-exome-sequencing; accessed on November 7, 2012.

THE CONTEXT-DEPENDENT CONCEPT OF "IDENTIFIABILITY"[18]

At the outset, Think Tank participants recognized that identifiability, the ability to link biospecimens and data to specific individuals, is a multifaceted concept. Think Tank discussions about the nature and meaning of "identifiability" led to the realization that the term "identifiable" is context dependent and, for this reason, potentially misunderstood or misused. It is helpful to distinguish different contexts in which the term "identifiability" has distinct and potentially contradictory meanings.

In medical research, biospecimens are identifiable if they are labeled with associated identifiable medical data pertaining to the individual from whom they are obtained, or if there are data or samples from a known individual to which the biospecimens may be linked. Key to the notion of identifiability in this context is the theoretical ability to re-identify an individual, or to determine private information about an individual, by matching de-identified biospecimens or genomic data to other identifiable information. Thus, access to de-identified biospecimens or data from a research participant will not, by itself, enable the donor to be identified. Additional information is needed —a matched biological sample, a medical record, demographic data, genomic or other -omics data, or some other independent information linkable to the donor. This is analogous to obtaining fingerprints at a crime scene; the information is only useful for identification if there are matching prints in a database of potential suspects enabling linkage to a known individual.

The potential accessibility of linkable data in today's highly networked culture creates an ethical conflict for the research community. On the one hand, we hope to expedite medical progress by broadly sharing research data. On the other hand, ethical considerations and regulatory constraints make us cautious about disclosing even aggregated -omics data if it is possible to trace the data back to individuals, particularly when the scope of prior consent for data sharing is unclear and re-contacting individuals is not possible or not contemplated.

Researchers believed initially that individuals would remain non-identifiable if research results were published at the group rather than individual level. This was accomplished by publishing allele frequencies of pooled datasets rather than individual genotypes. However, the paradigm-shifting work of Nils Homer et al.[19] showed that even trace amounts of genomic DNA of specific individuals within complex pooled genomic mixtures could be detected based on statistical analysis of allele frequencies of thousands of polymorphic genetic markers. Thus, as some think tank participants cautioned, with as little genetic material as the DNA obtained from a discarded paper coffee cup it could theoretically be possible to use Genome Wide Association Studies (GWAS) datasets to associate particular individuals with specific diseases.[20] How realistic this

[18] For further discussion of identifiability in the genomic era, see: Lowrance WW and Collins FS. 2007. Identifiability in genomic research. *Science* 317 (5838):600-2.
[19] Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 4(8):e1000167.
[20] Editorial Note: The President's Commission for the Study of Bioethical Issues, in its recent report Privacy and Progress in Whole Genome Sequencing (http://www.bioethics.gov/cms/node/764), notes that it is legal in many states to furtively send

scenario is and how to assess the resulting risk of harm was a subject of intense debate during the workshop. The challenge for policymakers, then, is to determine the extent to which data sharing policies and data security practices should account for such potential re-identification and any potential harm from it.

Despite the possibility of re-identifying individuals from coded and even pooled genomic data, federal regulatory approaches to identifiability consider de-identification of individual data a virtual safety net. Under HIPAA, for example, individual data can be accessed and shared without obtaining authorization from research participants if identifiers specified in the Privacy Rule of HIPAA are removed.[21] Under the Common Rule, if an individual's identity cannot "readily be ascertained or associated" by investigators conducting research with existing biospecimens or genomic data, then the research does not meet the regulatory definition of "human subject"[22] or require IRB review or informed consent, despite the fact that it may be possible to re-identify a live individual.

Hence, under current federal regulatory policy, if research activities involve randomly coded genomic data,[23] and the code is not available to the researchers holding the data, then the Common Rule considers individual identity not readily ascertainable, and HIPAA considers the data de-identified, even though someone may theoretically be able to identify an individual research participant by matching the coded data with other identifiable data.

Given the various uses and meanings of "identifiable" data versus "de-identified" data in research and regulatory contexts, deliberations about identifiability could be advanced if greater clarity regarding context were introduced into discussions about identifiability.

someone's discarded saliva sample to a commercial sequencing entity without that person's knowledge or permission and for malicious purposes, for example in contested custody other hostile litigation contexts.

[21] http://privacyruleandresearch.nih.gov/pr_08.asp#8a.

[22] 45 CFR 46.102(f).

[23] In accordance with guidance from the HHS Office for Human Research Protections (OHRP), the entity that oversees Common Rule implementation, downstream investigators working with coded biospecimens and/or genomic data are not working with "human subject" material as defined in the Common Rule if they do not have access to the key to the code and agree to not seek such access. See http://www.hhs.gov/ohrp/policy/cdebiol.html. Thus, investigators working with coded biospecimens and data that could potentially be re-identified do not need to obtain IRB review or seek informed consent for the research.

EXECUTIVE SUMMARY

On June 11 and 12, 2012, the National Cancer Institute (NCI) convened a Think Tank on Identifiability of Genomic Data and Biospecimens. Laura Buccini, Ph.D., of the Cleveland Clinic and Case Western Reserve University, and Carol Weil, J.D., of the National Cancer Institute, served as co-chairs. The purpose of this Think Tank was to bring together a multidisciplinary group of experts to explore challenges surrounding the identifiability of genomic data and biospecimens. The resulting considerations regarding consent policy development, data access and security policies, proposed best practices, and opportunities for empirical research, are summarized in this report.

Genomic research has moved quickly from genome-wide association studies in 2005, providing hundreds of thousands of data points per individual, to whole genome and exome sequencing, which can now reveal de-novo mutations that may be unique to a single human being. Economic projections indicate that these technologies will become affordable to large parts of the general population in the foreseeable future. This rapid progress in genomics and the dramatic increase in the amount of data generated raise questions about risks and benefits to research participants and the adequacy of existing rules intended to protect their privacy and confidentiality. Current federal regulations have been criticized for applying a binary approach that does not adequately reflect the nature and risks of the research; current federal standards render privacy protection unnecessary for genomic data that are stripped of identifiers. Research using de-identified genomic data is neither subject to the Common Rule protecting human research participants[24] nor to the HIPAA Privacy Rule.[25] Recently published studies suggest that biospecimens and derived data from pooled "de-identified" samples can be rendered identifiable if a matched sample is available. Concurrently, increased use of social networking sites in combination with companies that offer genomic data directly to the consumer increases the likelihood that individuals will share their information publicly. This raises new questions about the balance between access to data and control over personal health information and privacy.

Over the past years, research participants and their communities have become more interested in collaborating with investigators to frame both the agenda for medical research and the environment for recruitment and enrollment. While privacy is still a concern, many patients and research participants are eager to share their medical information for public benefit. Such "information altruists"[26] seem less fearful about the potential misuse of their medical data and will donate their data for future research uses even when there is no promised benefit for themselves or loved ones. More research is needed to explore the underpinnings of information altruism, its relative impact across disease and population communities, and the extent to which information altruism is motivated by trust for specific researchers rather than belief in the value of research generally.

[24] Code of Federal Regulations, Title 45, Part 46.
[25] http://privacyruleandresearch.nih.gov/pr_04.asp.
[26] Kohane IS, Altman RB. 2005. Health-information altruists--a potentially critical resource. *N Engl J Med*. 353(19):2074-7.

Beyond the phenomenon of information altruism, the development and growth of Internet technologies for both commercial and social exchange has made many individuals increasingly comfortable with the idea that their personal information can be accessed electronically by a broader segment of the public. The explosive popularity of communication tools such as Facebook[27] and Twitter[28] may reflect a cultural shift toward greater acceptance of open source data sharing. In this environment, it seems plausible that public interest in sharing research data might increase if we could effectively communicate the potential to advance medical progress through widespread "-omics" data sharing.

Irrespective of their personal views about data sharing and privacy risk, all research participants are entitled to respect, transparency, and choice regarding who can access their specimens and data (such as basic research institutions, educational institutions, the patient advocate community, or for-profit companies). Furthermore research participants must be offered the right to discontinue participation in research by withdrawing their consent for use of biospecimens and data in future research. Such decisions, while uncommon, are potentially influenced by sensationalist headlines and controversies. Notorious examples include the Havasupai lawsuit alleging misuse of biospecimens by secondary researchers; the furor over nonconsented storage of newborn blood spots for public health research; and the widely read story of Henrietta Lacks, a cancer patient in the 1950s whose remnant tissue, unbeknownst to her, was developed into cell lines for extensive future research. All three of these narratives highlight shortcomings in transparency and communication of respect for the contributions of research participants. It seems likely that such cases decrease individuals' willingness to participate in research and may breed resentment and fear, leading ultimately to a loss of public trust.

In order to counteract this trend, researchers must not only understand and appreciate the ethical parameters for conducting genomic research with biospecimens, but also engage the participant community on these issues. Participants at the think tank discussed several possible strategies for incorporating principles from patient- and community-centered initiatives in future research practice and policy. As an example of a topic deserving of more public input, DNA sequencing provides considerable information about the individual being sequenced, but potentially even more information about that person's offspring and unborn generations to come, because our capability for understanding the data will increase in future years. An individual's decision to release genomic information can therefore involve unpredictable informational risks for "unknowable" loved ones.

Another focus of the think tank was the recent Advance Notice for Proposed Rulemaking (ANPRM) to revise the federal regulations protecting human research participants, issued by the U.S. Department of Health and Human Services on July 22, 2011.[29] The ANPRM solicited public comment on specific questions related to appropriate federal protections for research participants, including protections for identifiable and de-identified biospecimens and data. Input was sought concerning the types of genomic data that should be considered identifiable, and the management of risks for inadvertent or intentional re-identification. The ANPRM defines a new risk category, "informational risk," which derives from inappropriate use or disclosure of

[27] https://www.facebook.com/.
[28] http://twitter.com/.
[29] http://www.gpo.gov/fdsys/pkg/FR-2011-07-26/pdf/2011-18792.pdf.

information that could harm a research participant or group, and it correlates risk with the nature of information and the degree of identifiability. The ANPRM proposes development of standardized data security protections that would remove informational risks from the purview of institutional review board (IRB) review, enabling researchers to make their own determinations. The ANPRM also contains a controversial proposal that would consider biospecimens identifiable in and of themselves, by virtue of the ability to obtain deoxyribonucleic acid (DNA) from them, even in the absence of any associated phenotypic or other data. Stakeholders in genomic research—including patients and other tissue donors, researchers, commercial biobanks, and policymakers—must consider the numerous changes proposed by the ANPRM because they would impact scientific inquiry, the autonomy and protection of research participants, and promotion of public trust.

Day 1 of the think tank began with welcoming remarks from Dr. Buccini and Ms. Weil. These were followed by a keynote session in which Dr. Laura Lyman Rodriguez, Ph.D., Director of the Office of Policy, Communications, and Education at the National Human Genome Research Institute (NHGRI), discussed evolving policy considerations regarding genomic data and identifiability, and Misha Angrist, Ph.D., Assistant Professor at the Institute for Genome Sciences and Policy at Duke University, discussed the perspectives of research participants. The rest of the first day was devoted to invited presentations.

Dr. Rodriguez began by providing historic context for the core issues of the current think tank. She noted that in 2006 the NHGRI sponsored a meeting[30] that led to an article[31] highlighting various definitions for "identifiable," the variety of means to render genomic data identifiable, and the individual differences in opinion, choices, and comfort level with respect to data-sharing and identifiability. Dr. Rodriguez further reviewed past and current data-sharing policies, and commented on the impact of recent high-throughput sequencing technologies, the benefits of which have recently been made available directly to the public by several private companies. She concluded her presentation by discussing the possible implications of the proposed changes in the ANPRM and emphasized the need for a proper balance between societal benefit and individual harm.

Dr. Angrist provided a unique account of his experience with issues of identifiability, combining his insights from being both an IRB member and a participant in the personal genome project (PGP). He contended that a lot of the problems with identifiability and privacy are directly caused by the fact that the research community and those overseeing protections of human research participants have lost sight of their intended audience. Dr. Angrist urged the participants to work toward a revision of the current system and cautioned that a dysfunctional status quo will worsen over time, given the rapid changes in technology, legislation, and infrastructure. He further noted that de-identification should not be used to relieve researchers from the obligation to interact with their research participants and concluded by highlighting the importance of the return of results for community engagement, transparency, and promotion of participant trust.

[30] The meeting website, which includes links to the full report, can be found at http://www.genome.gov/19519197.
[31] Lowrance WW and Collins FS. 2007. Identifiability in genomic research. *Science* 317 (5838):600-2.

Bradley Malin, Ph.D., M.S., of the Health Information Privacy Laboratory at Vanderbilt University, noted that the risk for re-identification is not unique to -omics data. Re-identification was demonstrated more than a decade ago by using demographic data, which are more widely available than genomic data. Dr. Malin further pointed out that even with a link such as a matching sample, re-identification, although possible, is not probable. He therefore suggested that privacy policy discussions should acknowledge that de-identification, while somewhat protective, is not a panacea. He further stated that privacy discussions should focus on the context of risk, opportunities for harm associated with re-identification, and risk-mitigation strategies.

Pearl O'Rourke, M.D., of Partners HealthCare, pointed out that the identifiability of various datasets or specimens lies on a spectrum, but that IRBs are forced to make a binary choice—yes or no—when determining whether a proposed study represents human subjects research. She added that all geneticists do not necessarily agree on whether particular types of -omics data are identifiable. Dr. O'Rourke explored the potential impact of the ANPRM, which would require a new brief (non-IRB) consent for all secondary research using biospecimens or genomic data.

Leslie Biesecker, M.D., of the NHGRI, discussed the current hypothesis-testing paradigm for clinical research and practice and proposed a different, hypothesis-generating paradigm. He noted that hypothesis-testing paradigms have been useful during decades of research using low-throughput technology, because these methods strictly limited the number of possible phenotypes under investigation. However, technology no longer constitutes the greatest bottleneck in genomics research, and Dr. Biesecker argued for a paradigm shift toward hypothesis-generating research, which allows investigators to assemble cohorts, collect molecular data, and use those data to identify new phenotypes and generate new hypotheses. Hypothesis-generating research could facilitate the transition toward personalized medicine and prevention, but it requires continuous interaction with research participants and an engaged, iterative approach to informed consent.

John Wilbanks, of the Ewing Marion Kauffman Foundation, pointed out that some individuals view the sharing of their own information as a form of control. In this manner, data sharing can be a beneficial way for sick patients to affirm their autonomy and overcome feelings of vulnerability and disempowerment. Unfortunately, there is currently no way for willing individuals to donate their health information to research studies. He described the development of Portable Legal Consent,[32] a transparent, digital informed consent process that explains the terms under which investigators can access personal data and the rights granted to researchers who access these data. During the process, participants receive a comprehensive explanation of the scope of their consent and information about the potential social and economic harms associated with open research and what happens if they choose to opt out. The Portable Legal Consent further facilitates communication between researchers and participants through a messaging system based on usernames.

Deborah Collyar, President of Patient Advocates in Research, provided a patient perspective and emphasized the need for plain language and clear information. She noted that in an age of patient-centered medicine, the identifiability of biospecimens and genomic data should be

[32] http://weconsent.us/.

discussed as it relates to the process of health and medicine and the effects on patients' care and quality of life. Consent documents should state clearly the risks associated with donation, address mistrust arising from past medical controversies or atrocities, and allow participants to decide whether they want information back. Ms. Collyar also noted that many patients want the biomedical research community to overcome intellectual property issues and share data.

Kenneth Chahine, Ph.D., J.D., Senior Vice President and General Manager of Ancestry DNA, noted that the world is changing from one in which health data are locked away in physicians' offices and research databases to one where individuals share large amounts of information through social networking and mobile phone applications. He pointed out that Ancestry DNA has a privacy portal but that many of its participants share their information to foster interactions. He also suggested that individuals' ideas about privacy could change when they see a personal benefit from sharing.

Jane Bambauer, J.D., of the University of Arizona, argued that the risk for re-identification is likely overstated because the demonstration attacks rely on special information or conditions that malicious intruders will not have. She cautioned against implementing protections when the credible downstream risks are still unknown and when the value of information sharing is great. Furthermore, the current consent model of privacy protection burdens patients with the responsibility to assess risks when regulators are in a better position to do so. With the assumption that the practical risk of re-identification is relatively small, Professor Bambauer proposed a three-step model that 1) includes a basic anonymization process, 2) holds data producers immune from privacy-related liability, and 3) ensures that malicious actors are held criminally liable for intentional re-identification and misuse of data.

Between the presentations on Day 1, think tank participants had opportunities to engage in moderated discussions, summaries of which have been integrated into the main part of this report in chronological order.

Day 2 was devoted to breakout discussions with groups focused on specific questions posed by the think tank organizers.[33] Breakout groups then summarized their discussions during plenary sessions. The think tank ended with a discussion of possible next steps and closing remarks from Dr. Buccini and Ms. Weil.

Think Tank Discussion Themes

Several themes arose during think tank discussions. It should be noted that these themes do not necessarily represent areas of consensus.

- All -omics data are theoretically identifiable, albeit at considerable effort and cost, and provided that a matching sample is available. However, data do not equate to knowledge or results, and the risk for re-identification from -omics data is still highly remote. Links to

[33] Members of the planning committee were as follows: Laura Buccini (the Cleveland Clinic), Mike Feolo (NCBI), Tiffany Green (NCI), Pritty Patel Joshi (OD), Christopher Kinsinger (NCI), Nicole Lockhart (NCI), Leah Mechanic (NCI), Stefanie Nelson (NCI), Laura Lyman Rodriguez (NHGRI), Kenna Shaw (NCI), Carol Weil (NCI), and Barbara Wold (NCI).

other datasets, or knowledge about an individual, are necessary to re-identify someone with confidence. Given these facts, consent form disclosures that minimize the risks of re-identification of stored genomic data remain appropriate today.

- "Identifiability" has been conflated with "privacy," which in turn has been conflated with "secrecy." In addition, the term "Identifiability" has different meanings in the realm of policy versus biotechnology. It influences our ability to generate data about research participants, the ability to return results to patients, and the ability to share those data in a way that promotes computational research. The concept of "identifiability" differs across those spectra, and the solution to one gap might widen another.
- Many patients or research participants are willing to share their data, and considerations of personal, familial, and societal benefit often outweigh individual privacy concerns. In some cases, patients or research participants view the sharing of their information as a form of control or empowerment. However, patients and research participants want more active involvement in the design and conduct of research, including the return of research results and a say over how their data or specimens are used. Quoting a patient advocate: "Nothing about us without us." A research system that respects participants' autonomy and rights creates public trust.
- Research using biospecimens or -omics data will be facilitated by a balance between potential liabilities to institutions and researchers and the need to respect the rights of participants and the community. Criminal penalties for non-research–related or malicious re-identification should be balanced with incentives for researchers who use -omics data appropriately. However, even if investigators who use potentially identifiable datasets in good faith are immune from privacy-related liability they should still be required to register for access to data, thus establishing an audit trail.
- Identifiability in the scientific context resides on a spectrum. When collected, biospecimens and data are naturally completely identifiable. Coded biospecimens and data are identifiable or re-identifiable, depending on the complexity of the coding system and the parties with access to the code key. DNA or RNA data stripped of identifiers but maintained in pooled databases may with effort be linked to matched samples, a potentially troubling form of re-identification. Despite this range of complexity in the research realm, current regulatory policy treats identifiability as a binary concept, triggering the application of human research protections for only identifiable biospecimens and information.
- Hypothesis-generating research, which has been made possible by high-throughput technologies, allows for broader assessments of phenotypes, but requires more interaction with research participants. This research paradigm, which involves greater community engagement, could facilitate the transition to personalized medicine, builds upon the interest and willingness of research participants to share and receive vast amounts of genetic and other private medical information.
- Social networking groups and commercial, direct-to-consumer companies such as 23andMe and Ancestry DNA might serve as models for community engagement in research. The audiences that utilize these services are, however, currently highly self-selected. Not everybody may be as curious about their -omics data, and little information is available regarding the attitude of the general population toward these technologies.

Suggestions from the Breakout Groups

- Several private organizations and companies are conducting studies that encourage people to share their -omics data. The Personal Genome Project (PGP), for example, aims to enroll 100,000 participants from the general public. Companies such as 23andMe, PatientsLikeMe, or Ancestry DNA provide mechanisms for their customers to share their data with other participants and/or researchers. However, no academic study has yet been conducted that aims at collecting an open-source dataset including a wide range of personal data. Such a project would be very valuable in order to systematically assess in an un-biased sample whether participants would consent to donate their data for research if they knew the dataset would be open and knew of the risks and benefits associated with data-sharing. Ideally the study would involve multiple subgroups stratified by comfort level, phenotype, demographic variables, and/or potential for stigma.
- Tweak, but do not redo or overhaul, existing National Institutes of Health (NIH)–supported databases, for example by introducing tiered consent structures that would allow a more refined distinction of different consent groups. Aim to maximize the integrity of data access for researchers while respecting the rights of research participants and engendering public trust.
- Consider new types of -omics data identifiable, but, as a default, give researchers controlled access to these data, for a wide range of research uses, until enough evidence warrants a reassessment of the risk of re-identification. Participants who contribute these data should be informed about identifiability and risk issues during the consent process.
- Streamline access to general controlled data through the establishment of fewer and better educated data access committees to address research requests. Such streamlining could facilitate researchers' access to a broad range of data.
- Incorporate transparency throughout the research process, not just when participants are giving consent to contribute specimens or data. Put more emphasis on community education and be skeptical of the amount of information that patients can understand in an acute, clinical situation. Participants should know what they are giving consent for, where their specimens might go, what researchers do or do not know, and how and whether results will be returned. Participants should be asked to provide and maintain their contact information if they want to receive research results.
- In addressing concerns about identifiability, assess what the greatest concerns of research participants are and categorize patients and research participants based on their privacy concerns. Understand that these categories are fluid, because participants' privacy concerns can change depending on the information they have about how use of their data improves public health.
- Incorporate evolving technology into current research study processes. For example, consent can now be obtained online, in a way that allows participants to choose how much information they want to receive.
- Encourage researchers to engage with policymakers to discuss law, policy, procedure, and practice around research using biospecimens and -omics data.
- More contentious were suggestions to cap liability for patients and participants in the event of data breaches, with some arguing to shift liability toward researchers and institutions and to impose a cost for data breaches.

Suggested Next Steps

- Conduct an empirical analysis of the risk or probability that a research participant could be impermissibly re-identified based on his or her -omics data.
    - This analysis should incorporate obligations and opportunities at various levels, including the consent process, custodianship and stewardship of stored and distributed data, and how to address the remote possibility of intentional misuse.
    - Results can guide the development of best practices for research institutions and biorepositories and can inform consent disclosures about privacy risk.
- Define and distinguish terms and concepts related to identifiability and risk. Consider the perspectives of and implications for the investigator, institution, community, regulatory entities, politicians, and the public. In particular, consider that "identifiability" is a loaded term and determine the most precise term to use. Precise definitions can aid policy developers by avoiding misunderstandings.
- Consider various policy options and the institutions that can respond to them. Policies can be implemented through the research industry, terms on National Institutes of Health (NIH) grants, endorsements from professional organizations, changes in regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the Common Rule, and legislative options.
- Develop workflow diagrams to assess the potential ramifications of deciding whether -omics data are identifiable. Such diagrams can illustrate the connections among definitions, laws, and policies related to privacy and identifiability, and they can turn an otherwise abstract discussion into a concrete one. Case studies incorporating changes suggested by think tank participants also could aid discussions.
- Emphasize that discussions of how best to address the identifiability of -omics data stems from an interest in the greater good and a respect for research participants' autonomy and rights. Acknowledge the size of datasets, the inefficiencies in moving data around, and the negative impact of these factors on research.
- Consider mechanisms to separate further consent to donate data or specimens for research from consent for a clinical procedure. One mechanism could resemble that for organ donation, in which individuals have time to consider risks or benefits and have a notation on their license if they are willing to donate data or specimens for research. The development of consent mechanisms also should consider who is administering consent.
- Foster collaborations between academic institutions, NIH, personal genome providers, advocacy groups, and other organizations.

THINK TANK SUMMARY

Introduction and Overview

The rapidly moving field of genomics can yield a wealth of information and facilitate the growth of personalized medicine. However, the rapid advancement also raises concerns about the risks to research participants associated with the fast-moving science, particularly to their privacy and confidentiality, and it is not clear that existing rules are adequate to protect against these risks. These concerns are illustrated by a recent Department of Health and Human Services (DHHS) Advance Notice for Proposed Rulemaking (ANPRM) for human subjects protections. The proposed new rule contains a provision implicitly specifying DNA as identifiable, even in the absence of associated data, phenotypes, or information from the medical record. Such a provision is complex, controversial, and multifaceted, raising questions regarding the ease of re-identification, the risks to research participants' privacy and the potential benefits of re-identification, the impacts of such a provision on policies governing investigators' access to specimens and data, and the ability to ensure the autonomy of biospecimen donors and preserve the public trust.

On June 11 and 12, 2012, the National Cancer Institute (NCI) convened a Think Tank on Identifiability of Genomic Data and Biospecimens. Co-chaired by Laura Buccini, PhD, of the Cleveland Clinic and Case Western Reserve University, and Carol Weil, JD, of the NCI, the think tank brought together a multidisciplinary group, including patient advocates, researchers, representatives of commercial genetic testing companies, NIH staff, and experts in privacy, bioethics, and health policy (see Appendix 1 for the list of participants). Presentations and moderated discussions highlighted challenges associated with defining the identifiability of genomic data and specimens and outlined considerations to inform policy development. Think tank participants also explored the efficacy of current data access and security policies, proposed best practices, and identified opportunities for empirical research. The think tank agenda is included as Appendix 2.

Following welcoming remarks from Ms. Weil and Dr. Buccini, keynote presentations were given by Dr. Laura Lyman Rodriguez, Director of the Office of Policy, Communications, and Education at the National Human Genome Research Institute (NHGRI), and Dr. Misha Angrist, Assistant Professor at the Institute for Genome Sciences and Policy at Duke University. The remainder of the first day involved presentations from invited speakers who explored the changing face of identifiability in the "-omics" era, the potential impact of the ANPRM on institutional review boards (IRBs), hypothesis-testing versus hypothesis-generating modes of research, new approaches to informed consent and human subjects protections, the future of genomic and health data, and legal safe harbors for research data. Think tank participants also engaged in moderated discussions throughout the day.

The second day of the Think Tank was devoted to breakout discussions, where groups considered specific questions posed by NCI:

1. What factors should be considered in the development of a Federal policy for access to publicly funded "-omics" research data?

- Are different types and models of access (open vs. controlled vs. hybrid) appropriate for different types and levels of data (individual vs. aggregate, genome-wide association studies [GWAS] vs. whole exome sequencing [WES] vs. whole genome sequencing [WGS])?
- Is there appropriate justification for treating "-omics" data differently from other types of research data?
- How should Federal policy take into account international data access and privacy standards?
- What research data or analysis is still needed to address these questions?

2. What considerations enter into determining whether -omics data are identifiable?
   - What distinguishes "identifiable" data from "de-identified" data?
   - Can data every truly be "de-identified," or is that concept outdated in the genomics era?
   - What criteria or standards should be used to establish whether particular types of -omics research produce identifiable or de-identified data?
   - What research data or analysis are still needed to address these questions?

3. What are the appropriate ethical constraints to allowing researchers broad access to -omics data?
   - What do we know about participant attitudes toward investigator access to their DNA and the privacy-utility tradeoff of limiting data access?
   - What do research participants and the public actually understand about the use of DNA in research (e.g., growth of cell lines, induced pluripotent stem cells), and what should they be informed about before consenting to participate?
   - To what extent should the concepts of autonomy, beneficence, and justice limit access by researchers to an individual's -omics data?
   - What research data or analysis are still needed to address these questions?

4. How can society minimize any risks and maximize any participant benefits of -omics research?
   - What are the risks of various -omics research technologies and the data they can produce?
   - How can -omics studies be designed to maximize individual participant, family, and community benefits (e.g., the return of individual or group population research results)?
   - What public or regulatory policies would promote appropriate balance of the risks and benefits of -omics research and help to avoid unwanted disclosures of identity and future uses of DNA for undesired purposes?
   - What research data or analysis is still needed to address these questions?

Following presentations from the breakout groups, the think tank closed with discussions of possible next steps and remarks from Dr. Buccini and Ms. Weil.

The remainder of this report summarizes the keynote and other invited presentations and accompanying discussion, as well as the report outs from the four breakout groups and associated discussion. Summaries of the four breakout group discussions are included as Appendices 3 to 6.

Is It or Isn't It? Evolving Policy Considerations Regarding Genomic Data and Identifiability

Laura Lyman Rodriguez, Ph.D., Office of Policy, Communications and Education, National Human Genome Research Institute, NIH

In 2006, prior to the broad availability of today's extremely data-dense -omics technologies, NHGRI held a meeting to explore technological, policy, and ethical issues related to the privacy, confidentiality, and identifiability of genomic data. At the time, Altman and colleagues had published an article[34] discussing the need for a balance between the level of privacy protection and the large amount of genomic information needed to advance science. The first draft of the HapMap (an international effort to establish maps of genetic variation in several populations, led by the NHGRI) had been completed and was beginning to be used to ask questions researchers had been unable to ask before. In addition, McGuire and Gibbs published an essay[35] suggesting that genomic data were no longer de-identified, and they called for the establishment of a new system for protections, such as a tiered consent model giving research participants a choice in how their data would be shared and used in the future. The 2006 meeting led to an article[36] that highlighted the various definitions for "identifiable"; the variety of means to render genomic data identifiable; the individual differences in opinion, choices, and comfort level with respect to data-sharing and identifiability; and the need to balance scientific potential with efforts to maintain public trust.

At the same time, NIH developed its data-sharing policies for GWAS, created the Database of Genotypes and Phenotypes (dbGaP), and established the policies through which it would operate. The regulations governing such research were—and are—based on the idea of identifiability: if data are not considered identifiable, then there is no requirement for IRB oversight or informed consent. Because data were de-identified during GWAS studies, such data were not considered identifiable and therefore not subject to IRB review. However, NIH established a policy that expected IRBs to review proposed GWAS for submission to dbGaP to determine whether data deposition and data sharing were consistent with the study's consent.

In 2012, analysis of molecular data has moved far beyond GWAS studies, giving researchers the ability to produce a massive amount of data quickly on a large number of individuals. With this technological capacity, perspectives about identifiability have shifted. Several researchers have shown that individuals' unique patterns in various genomic data types could be resolved within pooled, public data if a matching sample is available.[37] Recently, additional work by Shadt et al.[38]

---

[34] Lin Z, Owen AB, and Altman RB. 2004. Genomic research and human subject privacy. *Science* 305 (5681):183.

[35] McGuire AL and Gibbs RA. 2006. No longer de-identified. *Science* 312 (5772):370-1.

[36] Lowrance WW and Collins FS. 2007. Identifiability in genomic research. *Science* 317 (5838):600-2.

[37] Craig DW, Goor RM, Wang Z, Paschall J, Ostell J, Feolo M, Sherry ST, Manolio TA. 2011. Assessing and managing risk when sharing aggregate genetic variant data. *Nat Rev Genetics* 12 (10) Sep 16:730-6.

Im HK, Gamazon ER, Nicolae DL, Cox NJ. 2012. On sharing quantitative trait GWAS results in an era of multiple -omics data and the limits of genomic privacy. *Am J Hum Genet* 90(4) Apr 6:591-8.

further shows that gene expression patterns (e.g., mRNA expression levels) can be used to infer individual genotypes with high enough certainty to link these patterns to individual DNAs and can thus serve as a proxy when genotypes are not available. In addition, direct-to-consumer access to genomic data and the technology to generate them has increased, raising additional concerns of privacy. It is no longer clear whether identifiability is the most appropriate factor for research oversight and data management. Nor is it clear whether and how risk to privacy might be calculated across the spectrum of -omics research fields, or how protections and oversight can be calibrated to differing levels of risk.

Some studies have found that concern about the privacy of their health and genetic information can influence research participants' decision making,[39] but these studies have been small and focused on hypothetical choices. The power of these results to predict what participants would do when presented with an actual research study is probably low. A very common and consistent theme, however, is that research participants want transparency, respect, and some manner of choice in how their information is used. Where privacy, confidentiality, and identifiability rank with respect to other concerns is a matter of debate. Many participants, although concerned about their privacy, also want to see the benefits of advancing research. Other data have also suggested that the level of monetary incentive that may be attached to research participation might also affect how much of a priority privacy concerns become in participants' decision making. More study is needed to clarify the varying degrees of perceptions and risk tolerance on the part of potential research participants.

Participants and participant/patient communities are becoming increasingly interested in and involved in efforts to set the research agenda and drive research forward. Groups are forming through social networks to share not only genomic information, but also information about their phenotypes and medications. Other groups are focused on ways individuals can control their privacy. Initiatives are under way to build large research commons, and companies providing personal genome information frequently engage in collaborative research programs. A recent paper[40] has outlined principles associated with participant-centered initiatives and proposed that the structure of academic research might have to change to account for them.

Fears about genomics have been exacerbated by mixed messages and confusing communication around genomics with research participants and the public at large. Even as *Time* called 23andMe's genome kit the Best Invention of 2008, many media headlines, which often are slanted more toward sensationalism than toward scientific accuracy, have stoked fear. As illustrated by the Havasupai tribe lawsuit, recent publications about Henrietta Lacks, and the

---

Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 4(8):e1000167.

[38] Schadt EE, Woo S, and Hao K. 2012. Bayesian method to predict individual SNP genotypes from gene expression data. *Nat Genet* 44:603-8.

[39] Oliver JM, Slashinski MJ, Wang T, Kelly PA, Hilsenbeck SG, McGuire AL. 2012. Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. *Public Health Genomics* 15(2):106-14. Epub 2011 Dec 30.

[40] Kaye J, Curren L, Anderson N, Edwards K, Fullerton S, et al. 2012. From patients to partners: participant-centric initiatives in biomedical research. *Nat Rev Genet* 13(5) May:371-6.

furor over newborn blood spots, the public's negative perceptions about genomics are fueled by not only privacy risks, but also a lack of transparency about how participants' data will be used. The Genetic Information Nondiscrimination Act (GINA), which provides protections in employment and health insurance, was signed into law in 2008. Although it received some publicity when it was enacted, there is a clear lack of education of the public about it, and fewer people know about the law today compared to then. Individuals need to be made aware of these protections, so that they may be encouraged to feel comfortable participating in research. In addition, further discussion is needed to clarify who should decide the balance between societal benefit and individual harm, as well as how to respect the wishes of individual participants and sustain the public's trust.

With regard to identifiability, biospecimens, and genetic data, the ANPRM attempts to create a risk-based system for oversight and management, adding an "informational risk" category in classifying identifiability, and requires written consent for the use of all research samples. As defined by the ANPRM, informational risk derives from inappropriate use or disclosures that could be harmful to a participant or group, and it correlates with the nature of the information and the degree of identifiability. However, filters and criteria are needed to guide how data are managed, and who determines and defines risk must be specified. Specific questions on how Federal regulations should manage risk for possible re-identification, whether to consider a human biospecimen identifiable, and whether and which types of genomic data should be considered identifiable are under discussion.

Can You See the Real Me? Human Patients and Human Research Participants

Misha Angrist, Ph.D., M.S., M.F.A., Duke Institute for Genome Sciences and Policy

With respect to protecting human research participants, Dr. Angrist contended that the current system needs to be re-examined, because the status quo no longer works. Technological, legislative, and infrastructural reasons explain why, if left unchanged, problems will worsen over time. However, according to Dr. Angrist, behind all these causes is the simple fact that the research community and those overseeing protections of human research participants have lost sight of their intended audience.

This issue can be illustrated by a recent experience of the IRB on which Dr. Angrist serves. This IRB was asked to review a biobank protocol to collect thousands of blood and tissue samples together with phenotypic information from patients undergoing surgery. The consent document for this protocol, which appeared to have been reviewed extensively by attorneys, said in essence that participants would receive no compensation, even though the biobank would attempt to make money from their samples; that the biobank could come back to participants for more samples and information; that participants would not be able to see their own data; and that if they were injured during their participation in this protocol, the institution would treat them, at the expense of the participants.

Dr. Angrist wrote a letter to the principal investigator (PI) and proposed the addition of a statement instructing participants to provide contact information if they wanted an annual notification summarizing which research groups were studying their samples and what the research topics were. The PI responded that such a proposal was not feasible because it would lead to patients badgering researchers. Even when Dr. Angrist suggested a communication in

which investigators' names and institutions were omitted and the biobank would remain the gatekeeper, the PI responded that the biobank is a disinterested third party and might not have a clear understanding of the research done on a particular patient's tissue. The PI further questioned whether the biobank would be required to talk with participants whose samples had not been used, and he was unsure how patient-specific communication would work. Dr. Angrist stated that he was not requesting a detailed understanding of what specific investigators were doing, that sentences from external investigators' IRB-approved protocols could be provided in plain language, and that personal acknowledgement would require more work but add value to the participants and ultimately strengthen the position of the PI and the repository. The PI has not yet responded to that last letter.

At a subsequent meeting of the IRB, Dr. Angrist raised the issue of personal communication. His ideas were met with concerns that allowing such communication would set a precedent and could bring research to a halt. Furthermore, it was argued that personal communication could violate HIPAA because patients would now be identifiable. The protocol was approved by a majority of the IRB despite concerns about the consent.

Dr. Angrist further noted that, from a patient perspective, the issue of genetic privacy is somewhat of a red herring. Genetic privacy is important, but concerns about it arise from a legacy of abuses perpetrated during the 20th century eugenics movement. The policy of keeping genetic data private at all costs makes no allowances for where the locus of control lies or whether that locus should change. Genetic privacy is seductive because HIPAA makes it so; scrubbing data and specimens of identifiers absolves investigators of the responsibility to communicate with research participants. However, if that ethos is allowed to continue, investigators and institutions will lose in a world where genomic sequencing is cheaper, information is portable, and more and more individuals are sharing their information through Facebook, 23andMe, Ancestry.com, and other sites. Traditional genetic privacy cannot assimilate a Facebook-driven world where participants have a wide array of choices.

For some patients and research participants, concerns about privacy could be outweighed by the potential benefits of research. In a study of 59 patients and caregivers,[41] all supported biomedical and data-driven research and considered it essential to the discovery of better treatments, while none had heard of the HIPAA privacy rule. When educated about this rule, however, the study participants became angry about its effect on research. As demonstrated by the active interest in medical research involvement and data sharing by individuals who participate in web-based organizations such as PatientsLikeMe and 23andMe, the benefits of personal access to genetic and health information likely outweigh privacy concerns for many people. Patients might be concerned about their privacy, but they might be more concerned about carrying a risk allele and not knowing about it because of de-identification, and while some might worry about the implications of such alleles on their health insurance premiums, others, such as Dr. Angrist, are more concerned with the implications of these alleles with respect to the risks their children face.

[41] National Health Council. http://www.nationalhealthcouncil.org/pages/page-content.php?pageid=142#Just_Released:_NHC_Focus_Group_Study_on_HIPAA_Privacy_Rule.

Although Evans and Rothschild have argued that return of results is not complicated,[42] it is a difficult undertaking that will not grow easier any time soon. Providing individuals with access to information about themselves, even if it is only research information, can be difficult and frightening, and it may turn out to be wrong. However, return of results is necessary for community engagement and for transparency and promotion of trust, and the research community must ask why technology and de-identification trump autonomy, particularly in a Facebook-embracing world.

Discussion Points: Can You See the Real Me?

Return of results and information. Although return of results furthers participant autonomy, complex research outcomes with pooled or individual data cannot easily be translated to useful clinical information for individuals. Moreover, it is difficult for most research participants to understand statistics or why preliminary research results of even the highest quality may not translate to the clinical realm, and may even be wrong. Vast amounts of the scholarly literature are incorrect, systems in place are already generating data that cannot be replicated, and information is only trusted by the field after undergoing additional credentialing processes. New processes might be needed, and a balance between respecting research participants' autonomy and protecting them from inaccurate information must be found. Patients should not simply receive a hard drive of all the information from a study, but they should have a say in what types of results are worthy of return. In addition, many patients already use PubMed and want to read the full articles, even if they may be unable to determine if some of the information may be wrong. Usually, they can improve their care simply by talking with their doctors about what they have read. Individuals should be empowered to exercise their right to access their data and control how their information is used. Finally, there is a spectrum of interactions between researchers and their participants, and not returning results does not have to be synonymous with never contacting the research participant again, for example to inform him or her about new studies that have been approved for his or her sample.

Considerations for returning results must also account for legal issues. For example, the Centers for Medicare and Medicaid Services (CMS) have not authorized returning results from a laboratory not certified under the Clinical Laboratory Improvement Amendments (CLIA). There should be some distinction between getting access to general information pertinent to one's health and obtaining individual results that might affect immediate decisions about the best care.

Risks associated with identifiability of genetic information. Several think tank participants used a credit card analogy to illustrate risk. Financial institutions generally manage large amounts of information and protect it successfully, and even though individuals do not publish their credit card numbers, they do tolerate enough risk to give their credit card numbers to untrained people and allow them to copy those numbers. However, customers faced with identity theft can change their credit card numbers, and financial protections are in place for customers whose financial privacy has been threatened or compromised. Genetic results, which provide a lot of information not only about the participant but also about his or her relatives, cannot be changed. However, GINA and similar state laws provide a measure of protection against

[42] Evans JP and Rothschild BB. 2012. Return of results: not that complicated? *Genetics Med* 14:358-60.

discrimination by health insurers and employers. Yet, unlike the credit card scenario, there are currently no protections to help when genetic information has been stolen and/or misused by members of the general public. Rather than restricting access for research, public policy might be better served by penalizing those who discover and use the medical data of others without permission and for illicit ends.

Discussions might also shift from emphasizing risk to emphasizing rights. There have been class-action lawsuits involving scenarios in which patients have submitted to genetic testing and their insurers use their genetic profiles to influence decisions about other business lines such as automobile or life insurance. Behavior-targeted marketing on the Internet shows users ads based on consumer usage patterns. Although this type of marketing has been touted as low-risk and might benefit users through targeted advertising it diminishes privacy rights. Life and health insurance continues to be a gambling game in which patients and consumers are worse off when companies have more information. Moreover, issues of stigma should not be ignored. Some phenotypes are associated with considerably more stigma than others, precluding any attempts to design "one size fits all" solutions.

The research enterprise faces political risks from horror stories of research gone too far. Dr. Angrist noted that many privacy laws arose from such horror stories, and accounts of wrongdoing are met by severe oversight arising from politicians' overreaction. In addition, several public identifications of political officials from open-access datasets have occurred in the past, as in the case of Robert Bork's video rental history or the re-identification of former Massachusetts governor William Weld from medical records. These cases highlight privacy risks and, although mainly done as proof-of-concept studies, may create fear that overshadows the potential benefits of genomics research.

Identifiability concerns as a shield. As pointed out by Dr. Angrist, concerns about identifiability might be used by the research community to avoid caring for research participants and to avoid communicating with the public about controversial research outcomes. In addition, the discomfort with sharing research results with participants might stem in part from researchers' discomfort with ambiguity, whereas patients might prefer ambiguous information to no information.

Benefits of sharing information. Discussions of privacy seldom acknowledge the potential benefit, both to individuals and communities, of sharing information. Some think tank participants argued that a continued focus on privacy risk will blunt the enormous potential of genomics, whereas an emphasis on the benefits of participatory data sharing could help in accruing the large numbers of participants needed to accelerate research in the genomic era. Because all humans are imperfect beings who carry recessive germline mutations, sharing information could democratize and thereby minimize these blemishes, perhaps reducing stigma. Under those conditions, more individuals might consent to broad cohort studies.

Respect and trust. Conversations about genomics research must focus as much on respect and trust as they do on risk and benefit. A lack of respect and trust, which involves context, cultures, and relationships, could partially explain the low enrollment in clinical trials. Research participants do not want to be called "subjects," and they do not want to be shut out of the

process. A change in culture will, for example, involve transparency by informing participants about the types of research that use the data or specimens they have contributed.

Identifiability of communities. Discussions of identifiability must also consider the identifiability of communities and potential group harms. Assessment of community risk and group harm involves qualitative and culture-specific considerations. Thus discussions of risk assessment should shift from an attempt to develop formulas and unilateral decisions by IRBs to engaging distinct communities in defining risk and determining how best to live with that risk. The environmental health sciences community addresses this concern by conducting outreach to communities and obtaining their buy-in to research. Genetics and genomics might benefit from adopting more of this approach to research involving humans.

Although the data security provisions requiring protection of all identifiable data in the ANPRM focus on the identifiability of genomic data, it might actually be easier to identify individuals using standard epidemiological measures and social data. In addition, many pharmaco-epidemiological studies have been done without consent or even notice. Thus, focusing on the identifiability of genomic data will miss other key concerns about identifiability in human subject research, and any permissions required to use genomic data should also be required for use of epidemiological data.

-Omics and the Changing Face of Identifiability

Bradley Malin, Ph.D., M.S., Health Information Privacy Laboratory, Vanderbilt University

In 2006 America Online (AOL) de-identified data for approximately 650,000 customers who had submitted a total of 20 million search queries over 3 months and placed these data on a publicly available system. Later that year, Barbaro and Zeller downloaded the dataset, identified an individual, and contacted her to ask if she had been aware that AOL had posted her data on a website. They published this story[43] in the *New York Times* in August 2006, and within weeks AOL took down the dataset and dismissed its research and project manager. The chief technology officer also resigned. A class action lawsuit was filed in September 2006. A similar case occurred in 2008, when Netflix posted de-identified data about movie uses for approximately 450,000 customers with a challenge to develop better prediction algorithms. Researchers at the University of Texas found that by cross-referencing the movies data with the Internet Movie Database, they could identify unique patterns and identify two individual customers. These events also led to a class action lawsuit, and like AOL, Netflix took this dataset offline. Both lawsuits have been settled.

-Omics data are highly dimensional, and as shown by Homer and colleagues, individual identification is possible through summary statistics.[44] Identifying an individual by looking at de-identified sequences alone is virtually impossible with current technology and without a matched sample or other associated information. However, if someone knows an individual's genomic

[43] Barbaro M and Zeller T. A face is exposed for AOL searcher #4417749. *New York Times,* August 9, 2006.

[44] Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* Aug 29;4(8):e1000167.

sequence, that person could conduct a statistical analysis to infer whether that individual's information is in a general population database, such as HapMap or in a case or control group in dbGaP. Like Homer and colleagues, researchers at the Indiana University were also able to identify individuals,[45] needing even fewer single nucleotide polymorphisms (SNPs) to do it. Likewise, Gitschier reported[46] that an adversary can use data on publicly available databases to identify a participant, and another study[47] showed that individual SNP genotypes could be inferred from genetic expression data.

Yet, as the AOL and Netflix incidents demonstrate, the re-identification problem is not unique to biomedical research and does not arise from the explosion in -omics data. More than a decade ago, Sweeney[48] was able to cross-reference hospital discharge data (available to researchers for a fee) with general demographic and voter registration data from publicly available sites to find a diagnosis for former Massachusetts Governor William Weld. In addition, 63 to 87 percent of the U.S. population is identifiable through demographic information such as birth date, zip code, and gender, and a series of studies and articles have suggested that few pieces of data are needed to uniquely identify an individual. Indeed, an article by Ohm in 2010 noted that re-identification has become easy,[49] pointing to a fundamental misunderstanding underlying discussions about privacy law and protections. Thus the DHHS ANPRM notes that one can potentially link DNA to otherwise available data to identify individuals and suggests that all biospecimens and research involving storage and secondary analysis should be broadly re-labeled as involving identifiable information.

However, even with linking mechanisms, the possibility of re-identification from biospecimens or -omics data does not mean that it is likely to happen, and demonstration projects do not equate to everyday events. Recently, DHHS issued a challenge in which it provided 15,000 de-identified records to an investigative team and asked the team to identify the individuals. The researchers purchased data from several sources, but, as reported[50] in 2011, they were able to correctly identify only 2 individuals out of the 15,000 records. Moreover, they were unable to say which

[45] Wang R, Wang X, Li Z, Tang H, Reiter MK, Dong Z. 2009. Privacy-preserving genomic computation through program specialization. *Proceedings of the 16th Association for Computing Machinery (ACM) Conference on Computer and Communications Security*, pp. 338-47.
[46] Gitschier J. 2009. Inferential genotyping of Y chromosomes in Mormon Founders and comparison to CEU samples in the HapMap Project. *Am J Hum Genet* 84:251-8.
[47] Schadt EE et al. 2012. Bayesian method to predict individual SNP genotypes from gene expression data. *Nature Genetics*, DOI: 10.1038/ng.2248.
[48] Sweeney L. 1997. Weaving technology and policy together to maintain confidentiality. *J Law Med Ethics* 25(2-3) Summer-Fall:98-110, 82.
Sweeney L. 2001. Computational Disclosure Control: A Primer on Data Privacy Protection. Unpublished PhD thesis, Massachusetts Institute of Technology. Available at http://dspace.mit.edu/bitstream/handle/1721.1/8589/49279409.pdf.
Malin B and Sweeney L. 2000. Determining the identifiability of DNA database entries. *Proc AMIA Symp*, pp. 537-41.
[49] Ohm P. 2010. Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA Law Review* 57:1701-77.
[50] Lafky D. 2009 (October 8). The safe harbor method of de-identification: An empirical test. Available at http://www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf.

two individuals they had identified correctly. Likewise El Emam and colleagues[51] conducted a systematic review of case trials and newspaper reports and found 14 known re-identification attacks on any type of data. Eleven of these were conducted by researchers as demonstration projects, only two studies followed a standard re-identification process, and for those that did use standard techniques, the rate of re-identification was low. Yet publication of these findings met with controversy among security researchers.

Privacy discussions should acknowledge that risk for re-identification can be defined and modeled in many ways and that risk depends on context, which includes the availability and replicability of data and the ability to distinguish individuals at the group level or the single-person level. Likewise, concerns about risks to privacy will differ across potential uses for re-identification. For example, concerns about implications for forensic scientists or paternity tests might differ from those related to uses by life science researchers. Although de-identification is not Fort Knox, it is a hurdle, and currently is providing a reasonable degree of protection. Thus stakeholders should identify the opportunities for harm, weigh the benefits versus harm, and consider risk mitigation strategies. Institutions should not make data public without knowing what the associated risks are, quantifying the risks if possible, and combining technological controls. Overall, stakeholders should not demonize de-identification, but they should acknowledge that it is not a panacea.

Discussion Points: -Omics and the Changing Face of Identifiability

Individuals are reasonable about risk except in instances where someone has been burned. For example, the Netflix and AOL lawsuits and the paper by Homer and colleagues all met with shock and overreaction. The Netflix and AOL lawsuits were settled, but it is not clear that the majority of users who had not self-identified would have been identified from these datasets. Individuals consider risks in contexts such as who their identifiable information might affect (i.e., the risk to me vs. the risk to my children) and the degree of transparency. It is possible that the AOL and Netflix datasets might have fared differently if users were told about them in advance and given the opportunity to opt out. Likewise, insurers should be transparent and give patients an opportunity to consent before selling their information to pharmaceutical companies for targeted marketing. When institutions establish their policies, they shape the perceptions of the public. Thus they should be transparent about the risks they understand and acknowledge that they will never know all possible risks. However, behavioral economics studies have shown consistently that when individuals consent, they heavily discount the risks, only to feel remorse or anger if something happens to them or their information.

Other Discussion Points

- Along with risk and benefit, consequences for unauthorized access of de-identified data also should be discussed. Misappropriation of other protected information is associated with a consequence. For example, identity thieves face criminal prosecution, and rogue individuals at an institution risk losing that institution's grant. However, there is no personal responsibility for individual researchers. Placing some onus on individual users of -omics

[51] El Emam K, Jonker E, Arbuckle L, Malin B. 2011. A systematic review of re-identification attacks on health data. *PLoS ONE* 6(12):e28071.

data, even through a simple registration procedure, could force individual researchers to become more aware of their responsibility when they access data.

- Individuals are usually more likely to diminish their privacy when they or their children have a serious or rare disease and sharing information is viewed as a benefit. These observations warrant further, systematic studies of Ethical, Legal and Social Implications (ELSI).
- When considering risks to privacy, stakeholders should consider the opportunity costs or risks of nondisclosure and of not using identifiable samples or data. The decision to share information is not just a personal preference, and the burden to know which knowledge may be helpful for others or the community must not be placed on the individual alone. Identifiable samples and data may be advantageous not only for communication with research participants but also for obtaining additional information that could advance research.

IRBs: Forced to Deal With "Identifiability" of Everything … a Daunting Mandate!

P. Pearl O'Rourke, M.D., Human Research Affairs, Partners HealthCare

The scope of oversight of clinical research is driven by the Common Rule definitions of "research" and "human subject." Of particular note is the definition of "human subject" as "a living individual about whom a researcher conducting research obtains data through intervention or interaction with the individual or identifiable information." IRBs are therefore charged with determining whether the information the researcher proposes to collect or access is identifiable. IRB members are not identifiability experts and must make do using the Common Rule, HIPAA, State laws, and institutional policies, all of which ignore the subtler realities of identifiability. At the end of the day, IRBs must impose a binary system of identifiable or not identifiable to each protocol. It is important to understand the regulations, guidance, and threats of regulation that the IRB must consider.

Dr. O'Rourke first discussed the requirements imposed by the Common Rule: The Common Rule allows a fair amount of judgment by stating that research data are identifiable if the identity of a participant can be readily ascertained by an investigator or associated with the data. Of note, the Common Rule also includes categories of research that are exempt from the regulations. Pertinent to this discussion is the category of research involving the collection of existing data, documents, records, and biospecimens if the sources are publicly available or recorded by the investigator in such a manner that the participant cannot be identified directly or through identifiers. In reality, identifiability is evolving constantly, and the answers to whether a particular research project uses or generates identifiable data lie on a spectrum. However, the Common Rule allows only two options: the research involves human participants (i.e., the data or specimens are identifiable), or it does not (i.e., the data are not identifiable).

Dr. O'Rourke then compared the HIPAA requirements to the Common Rule: HIPAA offers two methods for determining identifiability. The first method states that data can be considered de-identified if the data are stripped of 18 identifiers. She noted that only 17 of these are specific, while the 18th identifier encompasses any other number, characteristic, or code that may be identifying. The second method requires that a person with appropriate knowledge and experience reviews the data and ascertains that they are not identifiable. HIPAA further includes the definition of a "limited dataset" that allows inclusion of limited identifiers (e.g., dates, geographic data) for specific uses under a formal Data Use Agreement. While many IRBs serve

as HIPAA Privacy Boards and must opine on HIPAA regulations, they must still apply the Common Rule in determining whether the data are identifiable and the research therefore is human subjects research. In addition to differences in determination of identifiability, the scope of HIPAA is different than the Common Rule. A comprehensive discussion is beyond the scope of this report, but a few differences are worth noting: HIPAA covers data of both live and deceased persons if that data are held by a covered entity (the Common Rule does not cover decreased persons) and there is no standard HIPAA implementation for research—in fact some research information is not covered by HIPAA.

Dr. O'Rourke went on to discuss how "genetic" data are considered according to these rules and regulations. Most IRBs have, so far, classified genomic data as unique but not adequate for identification without a link. However, even with allegedly anonymized data, IRBs are now considering whether enough independent identifiable genetic analyses exist to allow someone to identify a supposedly anonymized research participant through a simple merge. IRBs rely on guidance and input from colleagues in genetics, but the message is mixed: some geneticists suggest that all genetic data be considered identifiable, whereas others maintain that re-identifiability, even if theoretically possible, could be costly and time consuming and would require very specific information technology (IT) expertise that is not readily available. Dr. O'Rourke noted that the ANPRM makes some of this discussion moot, because it proposes that all tissue be considered identifiable. The importance of the determination of identifiability is best presented by following how tissue gets into research.

Figure 1 is a flow chart that captures the actions that must currently be taken whenever tissue is used in research. Separate pathways are created for tissue that is obtained solely for clinical care versus tissue that is obtained specifically for research purposes. For tissue obtained for clinical care, if there is tissue remaining after all clinical uses (including requirements for storing pathology samples) have been completed, then this tissue may be used in research. For this category, if remaining tissue is de-identified, then its use for research purposes does not constitute "human subject" research; whereas, if it remains identifiable, then its use in research does constitute human subject research and an IRB must review and determine if consent is required. For tissue initially obtained specifically for research, there must be an IRB approval and informed consent for the collection. As illustrated in the flow chart, if this research tissue is then going to be used for secondary research, then further oversight by an IRB will depend on whether the tissue to be used for secondary research purposes is identifiable. If it is identifiable, then IRB review is required. If the secondary researchers do not have access to identifiers, then the tissue is considered de-identified for purposes of the secondary research. The secondary research is further considered not to involve any "human subject" and does not require IRB review. Note that in many institutions, even though the human protection regulations do not require it, IRBs check to be certain that any secondary research use is consistent with the initial informed consent regardless of the identifiability of the tissue.[52]

[52] All figures are reproduced herein with permission from the author.

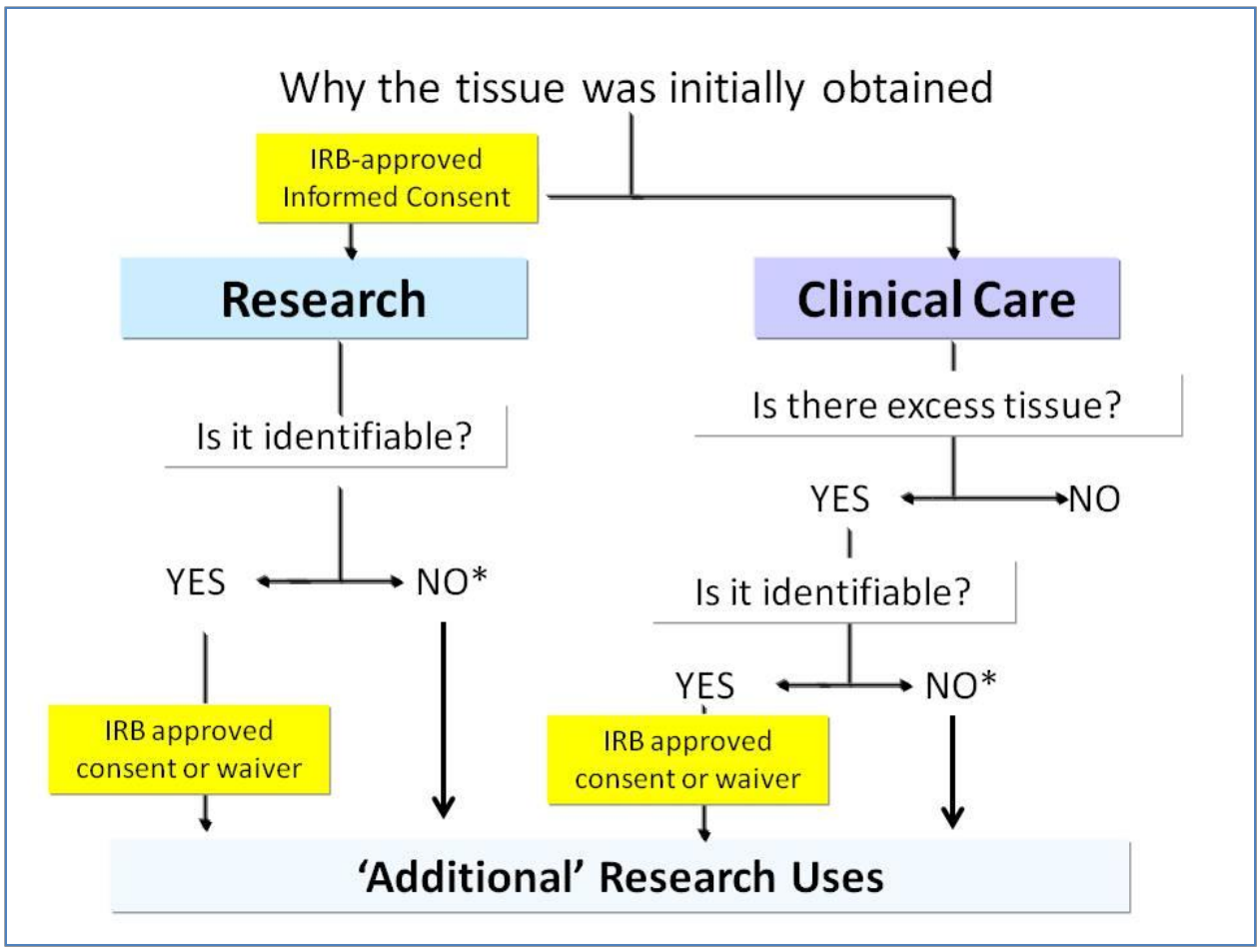

Figure 1: The status quo

Figure 2 is the same model diagram showing the decision tree if the current regulations remain, but all tissue is considered identifiable. In this scenario, there must be an IRB action for the use of any tissue in research—because the presumed identifiable status renders all research involving human tissue "human subject" research.

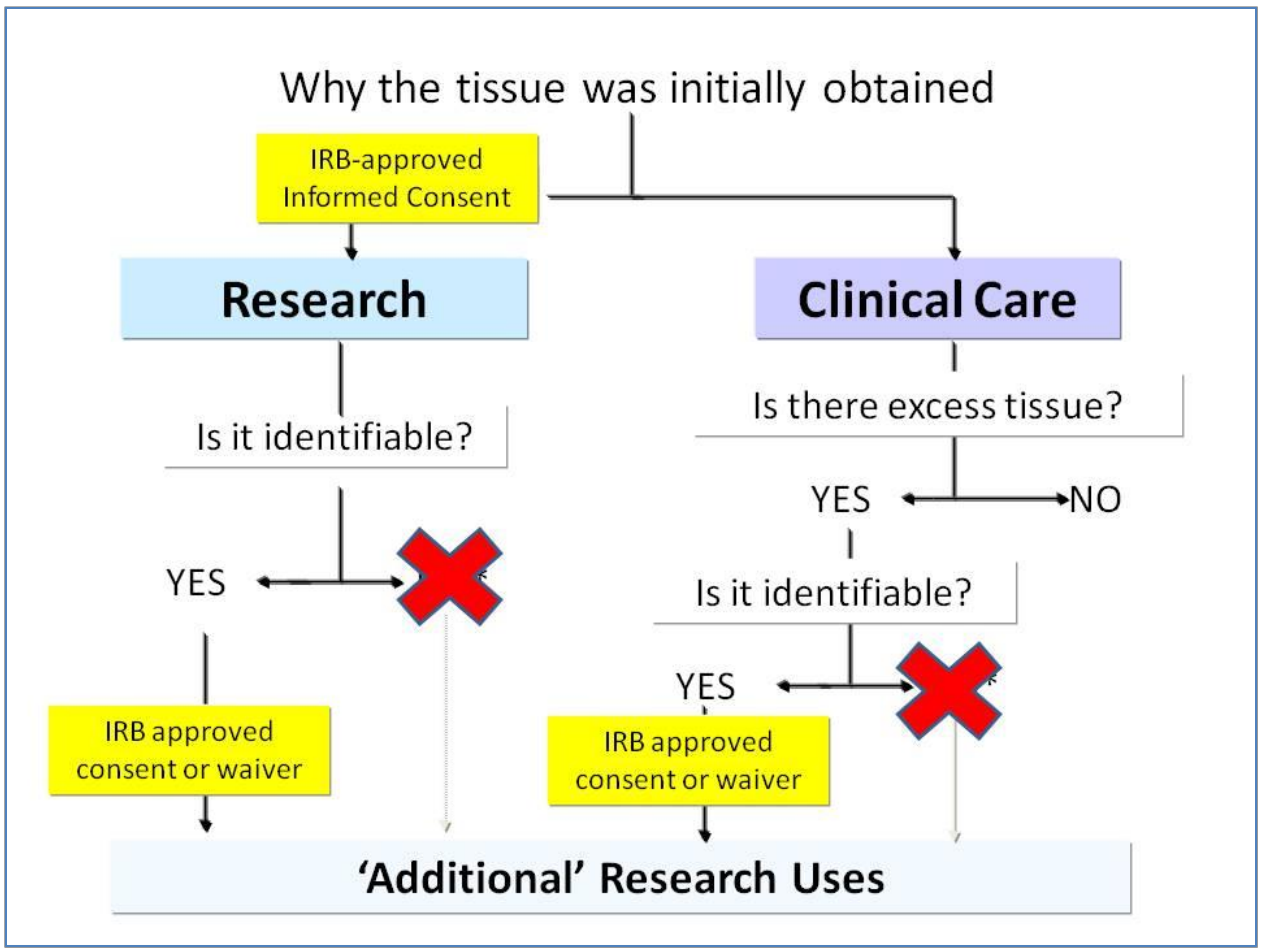

Figure 2: Current regulations—but all tissue considered identifiable

Figures 3 and 4 illustrate the effect of the ANPRM. Figure 3 addresses the changes that would be in place for tissue that is initially obtained for clinical purposes. The ANPRM would require a new "brief, general" consent that all patients would sign upon entry into the health care delivery system. This consent would address the use of that person's data and tissue for future research. Tissue and data from individuals who did NOT sign this new consent could NOT be used for research purposes. This brief consent would cover all downstream uses of clinically obtained data and tissue, perhaps with some disallowances for specified controversial uses such as reproductive research, and no IRB review would be required.

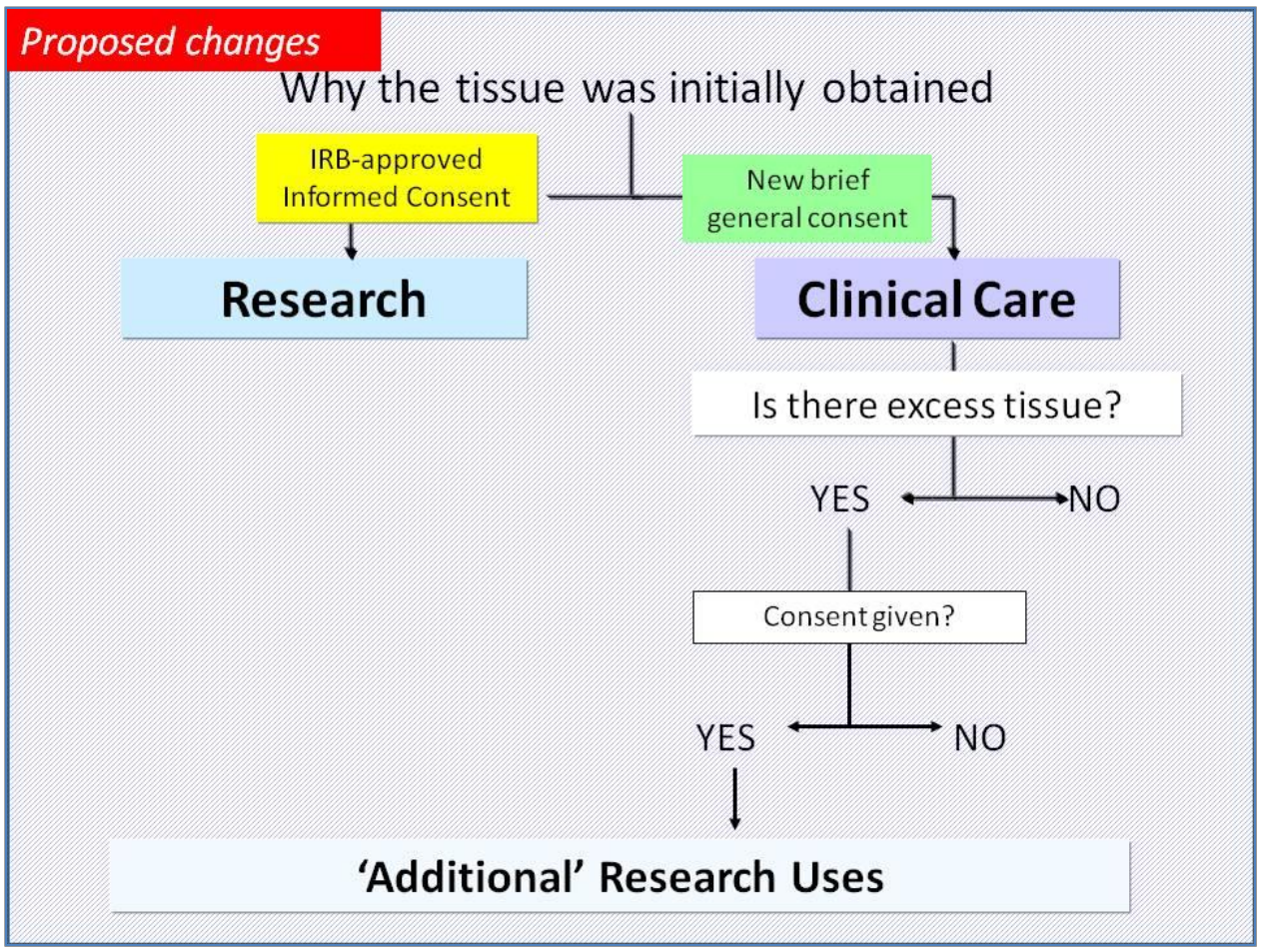

Figure 3: The effect of the ANPRM on research uses of clinical tissue

Figure 4 addresses the impact of the ANPRM for tissue obtained for research rather than clinical purposes. When tissue is initially collected for research purposes, IRB review and informed consent will be required as is the current practice. In order to use existing tissue (identifiable or de-identified) for a secondary research purpose, the brief ANPRM-proposed consent form is required. It is possible that the initial IRB-informed consent could contain the details of this new brief consent. But, if not, there would have to be a process by which the new short consent is obtained prior to secondary research uses.

Hence, the ANPRM decreases the role of IRB oversight for research involving human tissue and imposes a new "consent" requirement that is outside of the IRB purview.

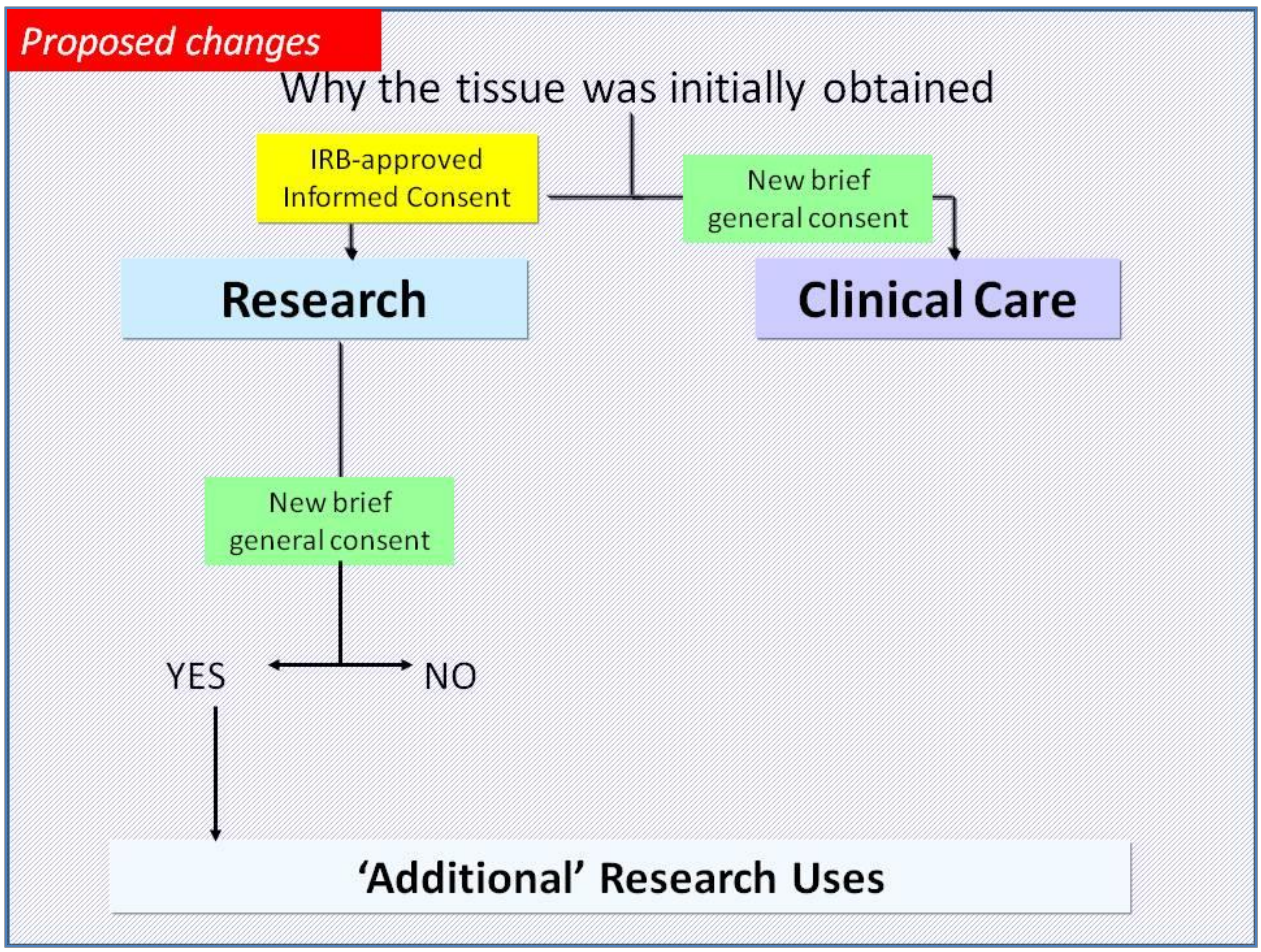

Figure 4: The effect of the ANPRM on tissue obtained for research purposes

Dr. O'Rourke concluded her presentation by noting that IRBs make decisions regarding identifiability on a daily basis. The current regulations rely on identifiability to decide what is/is not within the scope of the regulations and hence under the purview of the IRB. As described above, the IRB looks to the Common Rule, the Privacy Rule, evolving informational technology, as well as scientific advances in genetics. But there is no single voice, and IRBs are forced to squeeze these evolving definitions into a binary system.

Discussion Points: IRBs and Identifiability

The Common Rule allows for judgment, but IRBs are local, interpret and implement the regulations differently, and are inconsistent regarding determination of identifiability. IRBs could be helped by the development of white papers or targeted guidance regarding identifiability. This guidance should be informed by geneticists, ethicists, regulators, and funding institutions, and would hopefully reduce conflicting interpretations by different IRBs.

- Additional discussion is needed to address the potential risks and harms associated with re-identification.
- The group noted that the progress of the ANPRM must be closely followed. Specifically, several components merit attention: 1) the proposal that all tissue be considered

identifiable, 2) the introduction of a new brief, general consent, and 3) the decreased role for IRB oversight.

- The concepts of identifiability and protection of privacy and confidentiality are being challenged by many social networking and social media outlets, including listserves and chat rooms for patients with particular diseases or conditions, where participating individuals willingly and in some cases eagerly share identifiable personal information.

Hypothesis-Testing and Hypothesis-Generating Modes of Research

Leslie Glenn Biesecker, M.D., Genetic Disease Research Branch, National Human Genome Research Institute, NIH

In the current clinical research paradigm, investigators gather background information, formulate hypotheses about attributes pertinent to disease, phenotype patients, gather samples from them, apply an assay, and interpret the data to test, refine, or extend their hypotheses. A similar paradigm works in clinical practice, with physicians formulating a differential diagnosis and using specific assays to test it. The hypothesis-testing model allows investigators to focus on their questions, particularly in a prospective way. They can limit their obligations to research participants, and their interactions with them can be focused, narrow, and unitary. With hypothesis-testing research, investigators can control variables; type 1 errors, where data are falsely associated, are seldom an issue. However, investigators have been forced to rely on low-throughput assays, which require them to specify the diseases they want to study very narrowly already at the very beginning of the study design. Over the past decades, these emerging genomic techniques have been rate- and time-limiting, expensive, and noisy. Hypothesis-testing research focused on a limited number of educated guesses regarding factors involved in the pathology, because investigators could not prospectively phenotype all human traits.

In work following the hypothesis-testing paradigm, Dr. Biesecker collaborated with others to study combined malonic and methylmalonic acidemia, a rare, recessive autosomal disorder. He and his colleagues sequenced the exomes of affected children and their parents. From among several genes identified as containing mutations that met certain criteria suggesting they may be pathogenic, a mutation in a gene called ACSF3 appeared to be the most likely cause of the disease. They then conducted a candidate gene analysis on other patients with similar phenotypes and found several rare variants of ACSF3. Furthermore, when they scanned the literature, they found an ACSF3 gene relative that was associated with a biochemical function that was similar to the metabolic defect found in humans. To estimate how many individuals carry the recessive variants, Dr. Biesecker turned to the ClinSeq™ cohort, which has enrolled patients affected to varying degrees by atherosclerosis and includes 572 exomes. ClinSeq is an open-ended study that allows for full sequencing and downstream iterative phenotyping. When Dr. Biesecker and colleagues examined the sequence data, however, they found one patient homozygous for recessive mutations in ACSF3 among "control patients" within the cohort. Under the old paradigm, under which individuals are rigidly assigned to one category (healthy control vs. patient), one would have had to conclude that mutations in ACSF3 did not cause malonica and methylmalonic acidemia,

However, this study was based on the ClinSeq cohort, which allows the researchers to re-contact and re-phenotype individuals. It was found that the mutation carrier was an older adult who had, indeed, dramatically elevated levels of methylmalonic acid and clearly experienced difficulties

related to it. Thus, a generally consented participant allowed investigators to explore her genome, and they found a previously unknown, milder expression of a disease that otherwise presents in childhood. Based on the iterative phenotyping, Dr. Biesecker and colleagues thus found that their initial understanding of the range and severity of disease manifestation was wrong.

Genomics research has solved the problems of throughput and rate and time limitations, while the noise problem has worsened. Genomics represents a new reality, and when the underlying assumptions of research change, investigators must change the way they do science. Dr. Biesecker reported from another study, in which his team followed up on incidental (i.e., "secondary") findings in a control cohort and cancer screening. The team members began with an initial survey of all genes and variants and then applied certain filters to identify variants with high predictive value of causing disease. They narrowed down their initial list to eight variants, five of which were clearly pathogenic. Almost half of participants carrying these variants had family histories of cancer. However, the other half had reported no family histories, but this may be due to lack of information rather than lack of genetic risk at least in some of these families. The investigators returned the results and managed patients based on their genomic data. Similar work has been done for other phenotypes, such as familial hypercholesterolemia and cardiomyopathy.

Thus hypothesis-generating clinical research can use -omics data to tailor hypotheses and tell investigators how and where to look for disease. Investigators can assemble cohorts, generate molecular data, and identify patterns and perturbations to generate hypotheses. They can then test those hypotheses and modify them based on iterative phenotyping. Such research requires ongoing interactions with research participants, as well as an engaged, iterative approach to informed consent. Hypothesis-generating research could facilitate prevention by revealing ways to assess susceptibility in otherwise healthy individuals. Moreover, hypothesis-generating clinical practice could find diseases no one had suspected and, for patients, end the diagnostic odyssey.

This does not mean that hypothesis-testing research should be eliminated entirely. This paradigm is useful for broad and basic studies to understand the pathophysiology of disease, and efforts to anonymize and protect data can be appropriate in many cases. However, hypothesis-testing paradigms are clearly problematic at the level of individual patients, and hypothesis-generating paradigms are needed to move from where we are to individualized medicine.

Discussion Points: Hypothesis-Testing Versus Hypothesis-Generating Research

Under the current paradigm, informed consent forms usually tell participants that research results will be returned if they reveal something that could be severe or life-threatening. Dr. Biesecker and colleagues are assessing how much utility participants derive from information about recessive alleles. In addition, how much individuals know or understand about susceptibility and their risk when they consent to participate in cohorts is variable, and there is always uncertainty even with -omics data. Thus, clinical judgment and knowledge are still needed in decisions about whether or how to return results. The mechanism of returning research results will depend on the implications of the results. Needless to say, life-threatening results would never simply be posted on a website.

Dr. Biesecker and his colleagues run CLIA-approved laboratories and have validated mutations in established genes such as breast cancer 1, early onset (*BRCA1*). However, they strictly interpret the CLIA rules, and because their sequencing pipeline is not CLIA approved, they do not return research results arising from this pipeline. The data Dr. Biesecker discussed have been deposited into the single nucleotide polymorphism database (dbSNP), and work is under way to deposit them into dbGaP. Participants have consented to have their data deposited into publicly accessible databases.

Other Discussion Points

- At present there is not enough infrastructure to support the submission of data, the return of results, and the leveraging of information to use it in clinical practice. The research enterprise is still designed to support low-throughput biology. However, there are more bioinformatics challenges and more opportunities to return results, and funders could explore new paradigms for distributing resources.
- Proteomics data from individual samples will allow researchers or practitioners to identify traits that patients already have, compared with genomics data that simply highlight susceptibility. Proteomics data could thus prove more challenging with respect to identifiability.
- In all consent procedures, human nature has to be taken into account. Humans are not rational decision-making machines; they tend to repress unwanted information and, depending on the outcome, may ultimately regret participation. There is evidence for differences based on gender regarding how people handle this process.

Portable Legal Consent

John T. Wilbanks, Ewing Marion Kauffman Foundation

Health data are increasingly easy to collect outside of the clinics. For example, from 1994 through 1996, investigators collected data on eating patterns and body composition for approximately 2,700 participants, at a cost of about $85 per person, and published their findings in 2007. Now, however, mobile apps allow consumers to photograph what they are eating and estimate how healthy they think their food is. One such app has collected almost 8.0 million ratings and 0.5 million meals from approximately 50 countries, all in 5 months, with no grant or IRB overseeing this collection. Only a minority of users, say 5 percent of the population, is likely willing to submit their data in this way. However, given the size of the mobile app market today,

the total number of participants is very large compared to numbers reached in current controlled clinical trials. Several similar apps are in the pipeline, all collecting data. Mass culture can move faster than institutions. For example, Kaggle has announced a contest challenging participants to re-identify individuals from 10,000 de-identified health records and find predictive models.

Some participants view the sharing of their own data as a way to maintain control, as demonstrated by the development of open-source software, Creative Commons licensed works, and the sharing of legal standards by nonprofit organizations. However, there is really no place for most individuals to share their personal health data voluntarily. Individuals can collect data and obtain their genotypes on their smartphones but cannot communicate that information compatibly without an intervention.

Metcalfe's Law states that the systematic value of compatibly communicating devices grows as a square of their number. Compatibility is usually achieved either by development of open standards, or by monolithic companies that singlehandedly determine the specifications.

Keeping information private may, therefore, hinder the benefits of compatible communication. For example, Merck has developed Synapse, a mathematical engine for identifying correlations, but when the company kept its data private, it could not get enough data inside the engine to make it work. Synapse has, since then, been converted into an open-access tool for all scientists, but it is still struggling to reach the number of data points that are required for the algorithms to make reasonable predictions. Future programs of this kind will, likely, be more advanced, and should be able to use diverse kinds of input, for example how much participants eat and move. Ideally, these diverse types of input will be synced to create perturbation profiles and drug signatures.

Mr. Wilbanks has designed Portable Legal Consent (PLC), a digital consenting process that allows willing participants to donate their information to science. The process is explicit and transparent, explaining the terms of use researchers must agree to before accessing data and the rights they are granting to researchers that use the data. Participants are informed that research using their information will be open, and they watch a video explaining the potential for economic or social harm. Participants also are informed that they can opt out but those data already distributed cannot be retracted if they opt out. After all these steps, participants indicate that they have read and understand the informed consent process, and they sign consent digitally. Once participants sign consent, they create a user profile with a unique username, and they can upload various types of data, including genome, genotypes, medical or health records, lifestyle information, and laboratory results. In addition, PLC has a messaging system that allows researchers to communicate with those usernames. Investigators can invite individuals to participate in a study or request to collect more data, without knowing the identities of the individuals they contact.

PLC has approval to conduct federated public recruitment, with partners such as the Genetic Alliance serving as recruiters. Recruitment was approved by the Western IRB in April 2012. Data are hosted by Mr. Wilbanks and Sage Bionetworks and can be syndicated to various computational research environments. So far, PLC has had a 50 percent dropout rate. This high rate may be explained by the extensive consent procedure applied by PLC, which actually makes

people aware of risks and consequences. The rate is expected to increase even further once the recruitment shifts from curious volunteers to a more general population.

Making Sense of Genomics While Protecting People

Deborah Collyar, Patient Advocates in Research

The great promise to develop targeted therapies based on biomarker research has, until now, only delivered a few successes, and each new treatment has had complications and issues. Researchers, clinicians, patients, and other stakeholders have learned that one institution or company alone cannot solve these problems. Yet the research system and competitive grant mechanisms still support and reward individual successes. In addition, even when success has been achieved, such as with drugs like Gleevec, treatment stops working in some individuals. In other cases, tests such as those assessing *HER2/neu* mutations are valid, but the biomedical community still does not know to what extent these tests will, ultimately, inform clinical decision making.

It is therefore important to remember that all data are not knowledge and that data can improve outcomes for patients only if the right steps are taken to transform data into knowledge. Transforming data prematurely or incorrectly has many consequences, including product development hampered by non-valid biomarkers, racial and ethnic minorities left out of the "medical miracles," clinical utility hindered by false positives or negatives, and rare cancers remaining unaddressed. All of these consequences lead to waste of time and money. They erode trust, and they cost lives.

DNA is a human barcode, and although genomic research offers the promise of "medical miracles," it is associated with a fear that everyone can know almost everything about someone and his or her family members. The question, then, is how to conduct research while protecting participants from the possible discrimination, stigma, distress, and financial and familial consequences associated with inappropriate access to and misuse of data. Most patients, particularly those affected by cancer or rare diseases, are willing to donate specimens and data, but they do not think about the consequences of donation. In addition, many medical controversies and atrocities have happened in the past. Consent documents must address these issues in plain language, and researchers must keep in mind that patients are contributors and end-users of data.

The transition to patient-centered medicine will require researchers and health professionals to employ a whole-body approach, rather than focus only on -omics data. Researchers and health professionals will have to consider how molecular pathways interact, how this interaction fits in with the process of patients' health and medicine, and how this process will affect patients' quality of life. As illustrated by Ms. Collyar's conversations with patients, decisions to donate specimens or data are influenced by how their donations will be used.

Patients usually understand that companies might derive commercial benefit from these donations at a later date, but many do not want to donate directly for commercial work. They are concerned about identifiability, but they also want to understand how identifiability affects other aspects of their care. They need to be educated about distinctions between "genetic" and "genomic," and they want information and test results that are clear and simple enough to guide

their decision making. Patients also understand that diagnostic and molecular information is updated as researchers learn more, and they want updates to be available to everyone. Furthermore, patients want a choice on whether and when to receive information, and consent forms should allow them that choice. If they elect to receive information, then they only want the information that is most important. To honor the true meaning of "patient-centered medicine," researchers and clinicians must thus communicate with patients in plain language, providing information at their level of understanding. Finally, researchers often study relative risks and risks to population, while patients mainly care about their absolute risk.

Researchers also should acknowledge that the level of "personalized medicine" has not yet been reached. However, medicine is fast approaching an era of targeted therapies. This era will require targeted patients, and researchers will have to determine how to identify these patients, how to interest them in donating specimens and data for research, and what prior messages must be overcome. Likewise, an era of targeted therapies will also require the inclusion of targeted partners, including the professionals and advocates seeing the patients, and an understanding of possible barriers. In addition, biomedical researchers will have to address the existing competitiveness associated with many sample collections and must establish repositories that act as truly independent or honest-broker resources. They also need to acknowledge publicly the existing lack of reproducibility, and resources should be allocated to addressing that gap. Finally, patients want the biomedical research community to overcome intellectual property issues and share biospecimens, results, and updates.

Discussion Points: Protecting People and Portable Legal Consent

Reactions to the PLC illustrate the desire of patients to participate not only as a subject of research but also as a user of the data. However, a healthy balance between the research community's willingness to share biospecimens and data and participants' doubts when faced with the consequences of that sharing is difficult to achieve. Possible solutions include consents that allow participants to opt out of sharing. In particular, research consents should inform participants clearly about the purpose of the research, and the consent process should involve an ongoing dialogue of information and disclosure. Lessons about consent and communication can be learned from existing registries. The consent process also must include an educational component that clearly informs participants of the risks and does not oversell possible benefits. In addition, balance between risk and benefit is needed not only for research participants, but also for PIs, institutions, and other stakeholders in biomedical research.

At present, PLC is working only with the Athena study, but Mr. Wilbanks is talking with investigators about other cohort studies. In addition, free documents and technology are available to any researcher who wants to collect data in this way, and work is under way on a plain-language postcard patients can take to their physicians. A Group Banking committee has created a plain-language informed consent document for future use of biospecimens. Group Banking also has created separate patient brochures and IRB education sheets. The NCI Cancer Therapy and Evaluation Program is currently revising its consent templates based on input from the Group Banking committee.

Crowd-sourcing is made possible by an increasing number of people who have access to their own data. Each new study can have its own rules on data sharing, but it must return data to

participants who want it, and technology can make it easy for individuals to access and share their data. Crowd-sourcing could facilitate the achievement of desired study sample sizes, more so than the current recruitment models that rely on PIs seeking competitive funding. At present, PLC aims to stay simple and patient focused. However, it could expand to implement a conformed or centralized IRB that allows institutions to work together. At a recent Institute of Medicine meeting, health insurance representatives stated a willingness to collaborate and have researchers use their data. Making data open and allowing them to be combined could make a large practical and financial difference.

Other Discussion Points

- It is possible that PLC will promote the organization of people who certify conforming implementations, leading to web standards for obtaining the required permissions for future uses of data in research.
- Genetic Alliance is working with Mr. Wilbanks to aggregate data to test PLC and to place data in a common repository. Another project aims to enroll patients with Parkinson's disease, perform a complete blood workup every 3 months for 2 years, and publish those data along with the patient's full sequence.
- The development of PLC involved a large investment in computational research to achieve buy-in from a large teaching hospital. Concerned more with returning data to patients, PLC is unrelated to the hospital's existing protocols. The IRB-approved consent for PLC informs participants that a large hospital wants to make it easier for them to donate data from clinical studies. The company hosting the data, rather than the teaching hospital, is held liable for any incurred damages.

The Future of Genomic and Health Data

Kenneth Chahine, Ph.D., J.D., Ancestry DNA, LLC

The medical world is changing from one in which records are locked in a physician's office to one where individuals share a large amount of information. Along with this shift is a large movement in health information technology, with investments in more than 60 digital companies collecting data. The majority of these companies are collecting data from electronic health records (EHRs) or from smartphone applications. The launch of Ancestry DNA involved a team of scientists and non-scientific web designers and thus struck a balance between communicating scientific information accurately and clearly. Research data used to be only accessible for individuals who could interact with local clinical trial centers. Now, companies such as Ancestry DNA and 23andMe provide easy access to large amounts of disaggregated data. In addition, there is an increasing focus on the end user and the ability to present all these data in a way that can be consumed and understood by the general public.

Ancestry DNA aims to be as transparent as possible and to keep its site simple, offering a place to answer its users' questions. With respect to data security, the company has undergone several internal and external penetration tests. Although the protections make it less likely that someone will hack the system, there is a larger risk for security breaches on the user's side, for example by using too simple passwords or posting relevant information on Facebook. In addition, security can be bypassed in the physical world: DNA can easily be acquired by picking up a cup after a person has used it. Although privacy is maintained, about 80 percent of users make their genetic

trees public and invite interactions. This is consistent with users participating at 23andMe, who also want to share their results.

In considering a balance between individual privacy and social good, the biomedical research community should consider what contributors receive back and how to make that result personal. Some participants express their privacy concerns when clinical trials are discussed in abstract terms, but if the conversation changes to a personal focus, then they might be more willing to donate their DNA.

Discussion Points: The Future of Genomic and Health Data

Patients, research participants, and consumers have a right to access and share their data, and research participants, who want a more active role in research, want to engage with researchers and share their information. Institutions thus have a responsibility to return data in a way that is accurate and understandable and educates them about what they see in their data and the risks they take. 23andMe views this engagement as a way to drive genetic literacy and facilitate techniques to obtain clinically actionable data. Crowd-sourcing is also beneficial, because it allows the community members to help each other and correct mistakes. However, it should be noted that companies such as Ancestry DNA facilitate population genetics; testing is not done in CLIA-certified laboratories.

Ancestry DNA makes a commitment such that consumers own their data and can share it with any database or resource that is capable of receiving this information. The company itself does not share data with other databases without explicit consent. 23andMe brings in participants with certain phenotypes and partners with other organizations on various projects. The company works in accordance with HIPAA and informs participants that their data will be released to particular partners. NIH cannot perform the same types of functions because data can only come from pre-organized, pre-approved studies, rather than from individuals themselves. In addition, unlike Ancestry DNA or 23andMe, NIH-supported databases do not have ongoing relationships with study participants. It is not necessary to manage all private and public databases the same way, but the research community should identify ways to integrate public and private databases for social good. With the desire for the benefits of sharing sometimes outweighing concerns about privacy, NIH could use private database models as a guide in interacting with participants and informing them of risk. Lessons should be shared among direct-to-consumer databases, publicly funded databases, and biobanks.

Several presentations have focused on patients or consumers who are willing to share data and seek out opportunities to participate in research. Others are anxious about privacy, the definition of which may vary widely between individuals. One think tank participant quoted a paper in the *University of Pennsylvania Law Review*, in which Solove[53] develops a taxonomy of privacy, a term which, depending on the context, can refer to anonymity, surveillance, and other concepts. Although the majority of the privacy debate is driven by Internet security rather than health

[53] Solove DJ. 2006. A taxonomy of privacy. *University of Pennsylvania Law Review* 154(3):477-560.

concerns,[54] many individuals are anxious about the possibility of being denied insurance. However, one may argue that these concerns are addressed by the system as it stands. Individuals who want to share their data but cannot because of privacy laws are often frustrated, because institutions protecting their privacy do not help them get healthy. Systems designed to allow individuals to share data and participate in research must be transparent and voluntary, and they must allow individuals to actively opt in to participation.

Efforts to build or use large datasets also must consider the diversity of participants not only with respect to risk tolerance, but also with ability to pay for services, socioeconomic status, racial and ethnic identity, and other factors. Most genetic research has been done in European populations. 23andMe has launched Roots Into the Future, an initiative that recruits African Americans in an effort to replicate known genetic associations or find new ones. Ancestry DNA has a large subset of African Americans, but its genetic information also breaks down European populations into subsets. It is possible that these data could illustrate a continuum of genetic variability, rather than concrete barriers.

Other Discussion Points

- Willingness to share data or biospecimens for future research is not static. The Framingham Heart Study re-consents its participants every year, and the composition changes each year for the 15 percent of participants who consent to have their data shared only for studies related to their disease. Decisions to restrict or open data sharing are based on personal experiences.
- Longitudinal research is labor intensive because of the need to stay in contact with participants, follow up, and keep them engaged. Returning data to participants who want them could address some of these problems by making their involvement more active.
- Transferring information is somewhat difficult because most health records are coded for diagnosis, treatment, and billing and not for outcomes. Although a lot of information in hospital records is not related to billing, the data are often unreliable for researchers.
- As illustrated by comments about DNA left behind on a cup, organizations can never implement enough protections to prohibit all theoretical possibilities to circumvent these protections.
- The barriers to recruiting participants to donate specimens and data may be lowered by making sure that the patients are rewarded by return of their own results.

Seeding the Data Commons: Legal Safe Harbors for Research Data

Jane Bambauer, J.D., James E. Rogers College of Law, University of Arizona

Public data, herein referred to as the Data Commons, have aided with information justice, replication, accountability, and the avoidance of biased research. The larger the public dataset, the more powerful it becomes to detect, for example, harmful side effects. Vioxx, which ultimately was taken off the market in 2004, would have been withdrawn much earlier if a large public dataset had been available. However, issues of privacy and access are still under debate.

[54] Solove DJ. 2008. The new vulnerability: data security and personal information. Chapter 6 in A Chander, L Gelman, and MJ Radin, eds. *Securing Privacy in the Internet Age*. Pp. 111-36. http://ssrn.com/abstract=583483.

Paul Ohm[55] argues that public data releases are no longer ethical because of the ease of re-identification, and he calls for privacy laws to be rewritten to eliminate dependence on anonymization and to restrict access to data. This point of view underestimates the value of public data. On the other hand, the Electronic Frontier Foundation and the Electronic Privacy Information Center call for consent to be required in almost every context, even for the release of de-identified data. Solove and Schwartz[56] propose that identifiable data should require notice upon collection and security once collected, but this proposal underestimates the possibility of unanticipated new uses.

Consent has become the gold standard for privacy, but this model for privacy can be problematic even when it protects patients. Consent-based models give patients a property right in the data or specimens they contribute. However, because they receive indirect benefits of health research, the argument can be made that their contribution to the Data Commons should be compulsory. In addition, patients often do not know how their data could be misused and thus are not the best people to be asked to identify predictable risks. Consent also introduces selection bias: racial and ethnic minorities and patients of low socioeconomic status are underrepresented in research benefit, but they are less likely to consent to donate their specimens or data.[57]

Professor Bambauer noted that the literature on de-identification has been somewhat misleading, often overstating the risk of re-identification. For example, the famous Sweeney study[58] re-identified Governor Weld by matching the gender, birth date, and zip code listed on his voter registration records to a de-identified hospital dataset. However, Sweeney assumed that, because Governor Weld's gender, birth date, and zip code were unique among voter registration records, he must have a unique match in the Cambridge, Massachusetts, population. This assumption turns out to be mistaken. The conditions of the Sweeney study were replicated by Barth-Jones,[59] who found 174 males in the U.S. census who shared Governor Weld's zip code and birth year. Thus, there was a 35 percent chance that the re-identification could have been wrong for these males. Of these 174 males, more than one-third were not registered to vote. These non-voters may have shared Governor Weld's birth date, illustrating that Governor Weld could not have been re-identified with confidence using the simple matching attack described by Sweeney.[60]

---

[55] Ohm P. 2010. Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA Law Review* 57:1701-77.

[56] Solove DJ and Schwartz P. 2011. *The PII Problem: Privacy and a New Concept of Personally Identifiable Information,* 86 N.Y. L. Rev. 18141814.

[57] El Emam et al. 2009. A globally optimal k-anonymity method for the de-identification of health data, Appendix A: systematic reviews on the impact of consent on health research. *Journal of the American Medical Informatics Association* 16(5):670-82.

[58] Sweeney L. 2001. Computational Disclosure Control: A Primer on Data Privacy Protection. Unpublished PhD thesis, Massachusetts Institute of Technology, available at http://dspace.mit.edu/bitstream/handle/1721.1/8589/49279409.pdf.

[59] Barth-Jones DC. 2012. *The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now*. Available at http://ssrn.com/abstract=2076397.

[60] The details about Governor Weld's medical treatment following his physical collapse at a graduation ceremony were reported in the national news, and this additional information allowed him to be re-identified without doubt.

However, based on an incomplete understanding of the Sweeney study, the Department of Health and Human Services concluded that 97 percent of individuals whose data appears in de-identified databases containing zip codes and birth dates could be re-identified with certainty.[61] The resulting overreaction profoundly affected the development of de-identification provisions within the 2003 HIPAA Privacy Rule, Safe Harbor, and data use agreement rules.

De-identification protects study participants to some extent. When data undergo basic anonymization procedures, the costs of successful re-identification are driven up, making such attacks less likely to occur. Meanwhile, the research benefits of widely available microdata are profound. Genomic data present unique challenges to de-identification efforts, but the realistic marginal risks are still remote. In order for research participants to be re-identified from de-identified genomic data, an intruder would have to have their individual genotypes in identifiable form. Realistically, the value of re-identifying an individual to discover phenotypic information may be marginal once an intruder already has the target's genotype and can exploit the genotype's predictive power; it is not clear whether re-identification would be worth the effort, or would add much to the already significant privacy concerns raised by the intruder's access to the participant's genotype.

Under these assumptions, Professor Bambauer proposed a policy that:

- Creates an easy-to-apply set of protocols for creating powerful obstacles for de-identifying a dataset.
- Provides data producers with immunity from privacy-related liability.
- Holds individuals criminally liable if they misuse the data to re-identify participants.

To create as significant a barrier as possible to re-identify a dataset, data producers would have to remove direct identifiers and use an unknown sampling frame or minimum subgroup count. Professor Bambauer urged policymakers to tackle the predicate, and more troubling, potential intrusion from an attacker's inappropriate genetic sequencing before saddling the research community with the burden of preventing theoretical *additional* privacy attacks on de-identified genetic data.

The suggested protocol could be implemented easily even by non-experts such as institutions that are not used to handling large research datasets. At present, only risk researchers have performed re-identifications, and the regulatory process should be deliberate and passive until enough is known about credible downstream risks. Thus, it might not make sense to regulate public use de-identified research databases when re-identification still appears to be a theoretical risk.

Discussion Points: Legal Safe Harbors

Researchers are increasingly engaged in international collaborations and thus using datasets developed outside the United States. Thus U.S. privacy rules should be harmonized with those elsewhere. However, patients in Europe and Asia might be less concerned about sharing their health information because of differences in community culture and access to health care.

[61] HHS Notice of Proposed Rulemaking for the HIPAA Privacy Rule. November 3, 1999.

It is not clear how the threat of legal sanction would work against international or off-shore intruders. If such an intruder re-identified data for a company, the company could be held liable for purchasing data it knows to be re-identified. Professor Bambauer' proposal would not be adequate if the risk for re-identification were significant. However, at present, the rewards for re-identification are likely not worth the effort. Medical identity theft is already criminal conduct, and other re-identification attacks could be addressed using existing tort causes of action such as intrusion upon seclusion. However, medical identity theft is typically carried out by fraudulent use of medical identification numbers; it is difficult to imagine a scheme involving the re-identification of genomic data.

The cursory anonymization that Professor Bambauer proposes might not work for millions of bytes of genomic data. In addition, the development of tools that facilitate dataset integration could increase the likelihood of correct re-identification. At that time, stakeholders will have to re-examine the utility of the de-identified dataset, the cost of opportunities foregone by restricting access, and the willingness of participants to accept risk. However, the downstream risks of re-identification, and thus the long-term adequacy of Professor Bambauer's proposal, are still not clear at present.

The same factors that make data or specimens more identifiable are also the most useful for computational research. It is not clear how the proposed policy would balance protections against identifiability with the need for useful datasets. An imperfect solution might involve licensure, rather than making de-identified datasets fully public. For example, dbGaP and the Interuniversity Consortium for Political and Social Research at the University of Michigan require users to identify and register themselves, thus providing a level of accountability. Such policies preserve openness while producing an audit trail.

About two-thirds of HIPAA Privacy Rule complaints are dismissed without investigation because the Rule does not apply. Enforcement of HIPAA has aimed to encourage voluntary compliance, but this does not seem to work. Instead, public frustration has reached a level that undermines science and the relationships between patients and physicians. Patients have more difficulty accessing their own records because of the assumptions and institutional risk surrounding HIPAA. Harmonization of HIPAA and the Common Rule are needed to address data access by various stakeholders.

Other Discussion Points

- The ability to police and criminally prosecute inappropriate use should not be symbolic. Although she envisioned a person being charged with a felony and imprisoned for 5 years for criminal re-identification, Professor Bambauer acknowledged that the need for criminal enforcement would prove her initial assumptions wrong.
- Conversations about privacy rules should also consider re-identification based on information obtained without breaching a dataset. Consumers will continue to share their own data, and once published, copies of these data will be available globally forever after. Shared data released by one individual may increase privacy risks for others who may be included in an aggregated dataset. Therefore, the autonomous choice of one person may have implications for the identifiability of many others. Personal accountability quickly becomes very complicated in this context.

Reports from Breakout Groups

Group 1

Group 1 proposed a pilot of a truly open-source dataset that would allow access to all types of data, including -omics data. Participants willing to contribute their biospecimens and data to such a dataset would be informed that all data would be made public and that anyone would be able to access them. They also would be informed up front about the risks and benefits associated with such a dataset. Ideally, the pilot would involve multiple datasets stratified by varying comfort levels, phenotypes, and/or potential for social stigma. However, in light of continued budgetary constraints, one focused pilot of an open dataset would be a good start.

The group also agreed that existing NIH-supported databases represent works in progress, with staff committed to maximizing the integrity of data access for researchers while weighing respect for research participants and engendering public trust. Therefore, the group suggested that NIH continue to tweak these datasets, but not to overhaul them.

Discussion Points

The group envisioned a dataset that stored all types of data, not just those derived from -omics efforts. Group members envisioned that such a pilot could explore the willingness of individuals to contribute data and biospecimens if they know the dataset will be open. They also suggested that such a project could provide a proof of principle and demonstrate the benefits of data-sharing versus the risk for re-identification, serving as a model to overcome privacy concerns. The pilot project would test real-world consequences of truly open-source data, would assess how many individuals enroll and what their experiences would be, and would provide empirical evidence about whether data and information would be shared more quickly. The pilot would be patient-centered and focus on "what real people think and do."

Other Discussion Points

- As understood by one group member, the pilot could attempt to assemble a large cohort of individuals with specific phenotypes and track it longitudinally, with broad consents to overcome issues of controlled access.
- Patient advocacy groups, public health departments, and epidemiologists could, among others, serve as partners in recruiting individuals for such a cohort. Academic centers were expected to be more hesitant.
- The majority of data continue to come from European populations. However, because of historical issues, it is not clear how to encourage more racial and ethnic minorities to contribute data or specimens.

Group 2

Group 2 agreed that all data are identifiable, but it also acknowledged that re-identification requires a link to a matched sample and that some data are at greater risk for identifiability than others. Although some barriers are necessary to prevent against abuses of data, many barriers restrict scientific discovery. The group therefore suggested that NIH consider the context, including who will have access, how they will use the data, where the data might go, and what

the risks are. New types of data should be considered identifiable, but that does not mean they should be stored in a locked vault.

The group focused on a three-tiered "Russian doll model," similar to that used by the National Center for Biotechnology Information (NCBI). In this model, the open-access tier (Tier 1) allows anyone to download data contributed by participants who have been informed of identifiability and risk issues but permit wide access. The general controlled access tier (Tier 2) allows investigators to apply for access to data that fall under Exemption 4 or data for which participants have consented to broad sharing. The restricted tier (Tier 3), which allows no data sharing, includes data from specific population studies in which consent has been given for one research study only. Group 2 noted that the tiers are based on consent, rather than the level of identifiability. However, the group suggested that new types of data should default to Tier 1 unless or until enough evidence warrants a reassessment of identifiability.

Group 2 also discussed a trusted partner model for resources that reside outside NIH and for managing Tier 2 data. Such a model would allow resources to serve specific populations who have specific data requests, with access controlled by an NIH access committee. However, a trusted partner model is more expensive and complex. In addition, archival longevity is at risk because of a lack of integration with other resources and because funding might not be sustainable. Group 2 suggested streamlining access to Tier 2 data by establishing one access committee to address all researcher requests. Such a process would facilitate researchers' access to a broad range of data.

Models of consent also were discussed. Some universities employ an opt-out model, in which contributors are informed about the potential benefits of allowing their data to be shared for research. When patients permit their data to be used for research, they often do not understand what will be done with their samples. However, they trust IRBs to ensure that research using these samples is ethical and appropriate. Researchers, databases, and biorepositories should build trust with their communities through consistent engagement, including veto power for communities and events during which study results are reported.

Privacy conversations focused on ways to modify liability with respect to identifiable data. Group 2 suggested capping liability for the patients or participants in the event of data breaches, similar to the liability caps for credit card users. The group also suggested imposing a cost for data breaches on the researcher and institution, criminalizing re-identification as medical malpractice, and underwriting an audit of dbGaP users.

Law, policy, procedure, and practice govern all of these discussions, and stakeholders in research using biospecimens and -omics data should engage in all four areas. For example, scientists can get more involved in policymaking and encourage lawmakers to respond to requests for information.

Group 2 also provided the following suggestions:

- Convene a task force to examine pragmatics and balance over time as the landscape of identifiability continues to shift.
- Hold broader discussions with other -omics communities.

Discussion Points

- During discussions regarding where to include new types of data, Group 2 focused on proteomics data, which at present are associated with low risk for re-identification because matched samples most likely do not exist. The group assumed that all types of data are identifiable, even when no known evidence supports that assumption.
- Privacy involves shared rights and responsibilities. However, individuals often have no privacy rights and no ability to protect themselves until major damage has occurred. Thus, when addressing data breaches or misuse, relying solely on liability is likely a mistake. The research community should continue to adopt elements of stewardship and good data practices.
- When discussing risk and community engagement, staff members at biorepositories and data repositories are most aware of political responses to mistakes. Community engagement must include the political community, which pays close attention to identifiability issues and has a large amount of influence.
- Group 2 acknowledged that some IRBs are already influenced by anticipated changes to policy and that relating HIPAA to the Common Rule remains a struggle.
- Discussions in Group 2 focused solely on digital data.

Group 3

Group 3 discussed the ethical constraints to allowing investigators broad access to data. It focused on the following points:

- Participants are willing to consent to use of their data, but that willingness stems partly from a desire to be involved and to know how their data are helping. Formal mechanisms are not needed to keep participants informed.
- Transparency should occur throughout the process, not just when participants are consenting to contribute specimens or data. Participants should know what they are consenting to, where their specimens might go, what researchers do or do not know, and how and whether results will be returned. However, the burden of contact and re-contact is not fully on the research community. Participants can be asked to maintain their correct contact information if they want to receive results.
- The group could not reach consensus on whether to shift from a focus on de-identification toward considering all data identifiable. Using identifiable data would facilitate respect for participants and the return of research results. However, such a focus also would limit some current practices and use of retrospective samples.
- The group also discussed the need to include new technologies into current research processes, for example obtaining consent online in a way that allows participants to choose how much information they want.

Group 3 also discussed the need to harmonize privacy practices with international and State privacy laws, engage with research participants to assess their feelings about the tradeoffs surrounding return of research results, track successes in reaching broad populations, and conduct case studies to identify and address problems. Group members also questioned whether participants in cohorts such as “Patients Like Me” may later develop concerns or second thoughts.

Discussion Points

Many researchers do not want to work with identifiable data. They view it as "nuclear" and might not fully trust the staff in their laboratory. At the same time, de-identified data are less costly based on current consent models. Moreover, participants could be given control over whether they receive information from the research, but institutions might not know how to handle these types of interactions. Some stakeholders emphasize removing as many identifiers as possible, whereas others view the inability to return research results as immoral.

Group 3 clarified its view that investigators should not be compelled to use identifiable data. Instead, regulators and funders should recognize a spectrum of research, some of which will require more identifiable data, and not try to apply one research model to all projects. Some think tank participants advocated not only for deterrents or punishments for potential bad actors, but also for incentives to encourage investigators who will handle identifiable data appropriately. There are likely multiple morally and ethically acceptable ways to design research, across continuums of protections and the amount of information to include. A broad, prospective consent document could accommodate multiple future uses and preserve diverse ways of asking research questions.

Other Discussion Points

- There is a tradeoff between transparency and open access. For example, if data are stored online and researchers do not know who is using them, there is no transparency for the research participant.
- Broad statements about returning research results should consider epidemiological results and new types of data.
- Clinical and research enterprises are starting to merge. Thus research records are no longer separate from clinical ones, and once research results appear in the medical record, insurers could access them.

Group 4

Identifying risk is challenging. All data are identifiable, provided that a matched sample is available, but the levels of risk vary across data types. In addition, although good intentions might fuel stakeholders' questions about identifiability, these stakeholders also might have varying conflicts of interest. Moreover, the risks for identifiability must be distinguished from the risk for harm—for example, identification of an adopted child's birth parents, risks to someone running for office, inability to obtain health or life insurance, risks to family members, or social stigma. Group 4 noted that real instances of misuse have related to issues of data security and that, so far, re-identification has occurred primarily within the research setting. Thus far, the risks to research participants remain remote.

Stakeholders in research using biospecimens or -omics data want to implement protections to engender public trust and maximize the potential for recruiting participants. However, the group agreed that efforts to address remote risks could minimize the benefits to research participants and the community and that studies should minimally report aggregate data. Group 4 also suggested using streamlined data access committees (DACs) to control the release of data and

requiring investigators to accept accountability by registering for access to data. In addition, training data are needed to develop and improve bioinformatics.

With respect to the regulatory perspective, Group 4 agreed that regulatory rules should be adaptable, but it focused its discussion on consent documents. The group discussed whether consents can be written to highlight unknown risk and whether regulations can be applied so they are not limited to genomics data. The group agreed on the need to explain, in plain language, the remote chances that participants will be re-identified. A common consent was suggested.

In response to questions about the research data or analyses needed to drive policy, Group 4 focused on the identification of potential and actual misuses. The group suggested assessing what research participants are most concerned about. It was suggested that patients and research participants could be classified into categories based on their privacy concerns. The level of privacy concern can change depending on the information participants have about how use of their data improves public health.

Discussion Points

The group, along with other think tank participants, emphasized that levels of data access should be based not on the identifiability of data, but on the level of participants' concerns about re-identification. Many advocated for informing research participants that investigators cannot guarantee against re-identification, focusing efforts on the participants who are most nervous, and identifying comfort zones with which investigators can work. For example, individuals might be sensitive about participation in a depression study, because of the stigma associated with depression. However, they might be less sensitive if depression is studied as a side effect of cancer treatment.

The semantics around identifiability are important. The groups agreed that all -omics data are identifiable, but only with a link. They emphasized this distinction because, in the current regulatory context, a blanket statement that all -omics data are identifiable will classify research using that data as human subjects research, triggering several burdensome consequences. However, promising that these data are not identifiable is also problematic, because of the journalistic and political risks associated with re-identification. In addition, it is not clear whether consent documents should continue to assure research participants of their privacy when increasing amounts of information are out in the open. Because the journalistic and political risk is highest before clinical value is demonstrated, the research enterprise should work quickly to show the public the clinical value of research using -omics data, rather than establish a policy that might hamper research. Investigators also should discuss absolute risk, noting that the risk for re-identification from genomic data is smaller than that for re-identification from medical records or even vehicle identification numbers. However, research stakeholders also must avoid using language that overhypes discoveries.

Suggested Next Steps

- Conduct an empirical analysis of the risk or probability that someone could be re-identified based on his or her -omics data. This analysis should incorporate obligations and

opportunities at various levels, including the consent process, stewardship of the data, and how to address misuse (although at present the probability of misuse remains remote).

- Define and distinguish terms and concepts related to identifiability and risk. In so doing, consider the perspectives of and implications for the investigator, institution, community, regulatory entities, politicians, and the public. In particular, consider that "identifiability" is a loaded term and find the most precise term to use. Precise definitions can aid those responsible for developing policy.
- Consider various policy options and the institutions that can respond to them. Policies can be implemented through the research industry, terms on NIH grants, emphases from professional organizations, changes in regulations such as HIPAA and the Common Rule, and legislative options.
- Develop workflow diagrams to assess the potential ramifications of deciding whether -omics data are or are not identifiable. Such diagrams can illustrate the connections among definitions, laws, and policies related to privacy and identifiability, and they can turn an otherwise abstract discussion into a concrete one. Case studies incorporating changes suggested by think tank participants could also aid discussions. Workflow diagrams and case studies should be incorporated into any white paper arising from this think tank.
- Emphasize that discussions of how best to address the identifiability of -omics data stem from an interest in the greater good and a respect for research participants' autonomy and rights.
- Consider mechanisms to separate further consent to donate data or specimens for research from consent for clinical procedures. One possible mechanism could be modeled on organ donation, in which individuals have time to consider risks or benefits and have a notation on their license if they are willing to donate data or specimens for research. The development of consent mechanisms also should consider who is administering consent.
- Contact private corporations that provide personal genomes to determine possible areas of collaboration. These companies usually have research arms and have expressed an interest in collaborating with academic institutions, NIH, and other organizations.

APPENDIX 1: PARTICIPANT ROSTER

*Member of Planning Committee

Misha Angrist, Ph.D., M.S., M.F.A.
Assistant Professor
Institute for Genome Sciences and Policy
Sanford School of Public Policy
Duke University
Box 90141
229 North Building
Durham, NC 27708-0141
(919) 684-2872
misha.angrist@duke.edu

Alice Bailey
Scientific Program Analyst
Office of Policy, Communications and Education
National Human Genome Research Institute
National Institutes of Health
Building 31, Room 4B-09
31 Center Drive
Bethesda, MD 20892
(301) 443-1847
baileyali@mail.nih.gov

Jane Bambauer, J.D.
Associate Professor of Law
James E. Rogers College of Law
University of Arizona
122 Smith Street, #2
Brooklyn, NY 11201
(718) 809-7030
janebambauer@email.arizona.edu

Leslie Glenn Biesecker, M.D.
Chief
Genetic Disease Research Branch
National Human Genome Research Institute
National Institutes of Health
Building 49, Room 4A-56
49 Convent Drive
Bethesda, MD 20892
(301) 402-2041
leslieb@helix.nih.gov

Ralph Bradshaw, Ph.D.
University of California, San Francisco
Genentech Hall, Room N472
Box 2240
600 16th Street
San Francisco, CA 94158-2240
(415) 476-3813
rablab@uci.edu

***Laura D. Buccini, D.P.H., M.P.H.**
Associate Staff and Associate Professor
Digestive Disease Institute
Transplant Center
Cleveland Clinic
Lerner College of Medicine
Case Western Reserve University
9500 Euclid Avenue
Cleveland, OH 44195
(216) 445-8125
buccinl@ccf.org

Kathleen Calzone, Ph.D.
Genetics Branch
Center for Cancer Research
National Cancer Institute
National Institutes of Health
Building 41, Room B622
MSC 5055
41 Medlars Drive
Bethesda, MD 20892
(301) 435-0538
calzonek@mail.nih.gov

Michele Cargill, Ph.D.
Chief Scientific Officer
Locus Development
458 Brannan Street
San Francisco, CA 94107
(925) 948-5213
michele.cargill@locusdev.net

Latarsha Carithers, Ph.D.
Project Manager
Office of Biorepositories and Biospecimen Research
National Cancer Institute
National Institutes of Health
11400 Rockville Pike, Suite 748
Bethesda, MD 20892
(301) 435-8437
latarsha.carithers@nih.gov

Kenneth Chahine, Ph.D., J.D.
Senior Vice President and General Manager
Ancestry DNA, LLC
360 West 4800 North
Provo, UT 84604
(801) 705-7014
kchahine@ancestry.com

Stephen Chanock, M.D.
Chief
Laboratory of Translational Genomics
Division of Cancer Epidemiology and Genetics
National Cancer Institute
National Institutes of Health
8717 Grovemont Circle
Bethesda, MD 20854-4605
(301) 435-7559
chanocks@mail.nih.gov

George Church, Ph.D.
Professor of Genetics
Department of Genetics
Harvard Medical School
NRB 238
77 Avenue Louis Pasteur
Boston, MA 02115
(617) 432-7562
gchurch@genetics.med.harvard.edu

Ellen Wright Clayton, M.D., J.D.
Craig-Weaver Chair in Pediatrics
Professor of Pediatrics
Professor of Law
Cofounder
Center for Biomedical Ethics and Society
Vanderbilt University
2525 West End Avenue, Suite 400
Nashville, TN 37203
(615) 322-1186
ellen.clayton@vanderbilt.edu

Deborah Collyar
President
Patient Advocates in Research
3687 Silver Oak Place
Danville, CA 94506
(925) 260-1006
deborah@tumortime.com

David W. Craig, Ph.D .
Deputy Director
Information Services
Director, Neurogenomics
The Translational Genomics Research Institute
445 North Fifth Street, Suite 500
Phoenix, AZ 85013
(602) 343-8767
dcraig@tgen.org

***Michael Feolo, M.S.**
Staff Scientist
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Building 45, Room 4AN.12B
Bethesda, MD 20894
(301) 402-2874
feolo@ncbi.nlm.nih.gov

Claudia Maria Gaffey, M.D.
Office of In Vitro Diagnostic Device Evaluation and Safety
Center for Devices and Radiological Health
U.S. Food and Drug Administration
White Oak 66, Room 5516
10903 New Hampshire Avenue
Silver Spring, MD 20993
(301) 796-6196
claudia.gaffey@fda.hhs.gov

Robert Gellman
419 Fifth Street, SE
Washington, DC 20003
(202) 543-7923
bob@bobgellman.com

Mark Gerstein, Ph.D.
Computational Biology Program
Molecular Biophysics and Biochemistry
Yale University
Bass 432A
266 Whitney Avenue
New Haven, CT 06520
(203) 432-6105
mark.gerstein@yale.edu

Elizabeth Gillanders, Ph.D.
Host Susceptibility Factors Branch
Epidemiology and Genetics Research Program
National Cancer Institute
National Institutes of Health
Executive Plaza North, Suite 5116
MSC 7393
6130 Executive Boulevard
Bethesda, MD 20892-7393
(301) 594-5868
lgilland@mail.nih.gov

***Tiffany Green, M.P.H.**
National Cancer Institute
National Institutes of Health
Executive Plaza North, Suite 5141A
6130 Executive Boulevard
Bethesda, MD 20892
(301) 594-7348
greentif@mail.nih.gov

Robert Grossman, Ph.D.
Chief Research Informatics Officer
Biological Sciences Division
Professor of Medicine
Section of Genetic Medicine
Director of Informatics
Institute for Genomics and Systems Biology
Senior Fellow, Computation Institute
University of Chicago
KCBD 10142
900 East 57th Street
Chicago, IL 60637
(773) 702-5660
robert.grossman@uchicago.edu

Sara Chandros Hull, Ph.D.
Director, Bioethics Core
National Human Genome Research Institute
Faculty, Department of Bioethics
Clinical Center
National Institutes of Health
Building 10, Room 1C-118
MSC 1156
10 Center Drive
Bethesda, MD 20892-1156
(301) 435-8712
shull@mail.nih.gov

***Pritty Joshi, Ph.D.**
American Association for the Advancement of Science (AAAS)
Science & Technology Policy Fellow
The Cancer Genome Atlas
Center for Cancer Genomics
National Cancer Institute
National Institutes of Health
Building 31, Room 3A-20
31 Center Drive
Bethesda, MD 20892
(301) 451-8767
pritty.joshi@nih.gov

Julie Kaneshiro, M.A.
Policy Team Leader
Office for Human Research Protections
U.S. Department of Health and Human Services
1101 Wootton Parkway, Suite 200
Rockville, MD 20852
(240) 453-8293
julie.kaneshiro@hhs.gov

***Christopher R. Kinsinger, Ph.D.**
Program Manager
Clinical Proteomics Tumor Analysis Consortium
National Cancer Institute
National Institutes of Health
Building 31
31 Center Drive
Bethesda, MD 20892
(301) 594-9016
kinsingc@mail.nih.gov

Joanne R. Less, Ph.D.
Director, Office of Good Clinical Practice
U.S. Food and Drug Administration
WO 32-5168
10903 New Hampshire Avenue
Silver Spring, MD 20993
(301) 796-8343
joanne.less@fda.hhs.gov

***Nicole Lockhart, Ph.D.**
National Cancer Institute
National Human Genome Research Institute
National Institutes of Health
5635 Fishers Lane, Suite 4076
Rockville, MD 20892
(301) 435-5697
lockhani@mail.nih.gov

Subha Madhavan, Ph.D.
Director
Clinical Research Informatics
Lombardi Comprehensive Cancer Center
Director, Biomedical Informatics
Georgetown-Howard Universities
NIH Clinical Translational Science Award (CTSA)
Georgetown University Medical Center
Suite 110
2115 Wisconsin Avenue, NW
Washington, DC 20007
(202) 687-3294
sm696@georgetown.edu

Bradley Malin, Ph.D., M.S.
Associate Professor of Biomedical Informatics and Computer Science
School of Engineering
School of Medicine
Director
Health Information Privacy Laboratory
Vanderbilt University
2525 West End Avenue, Suite 600
Nashville, TN 37203
(615) 343-9096
b.malin@vanderbilt.edu

Stephen Vincent May, M.A., M.S.W.
Executive Director
Forum on Genetic Equity
38 Whitehead Avenue
Hull, MA 02045
(781) 724-2921
fge_steve@yahoo.com

***Leah Mechanic, Ph.D., M.P.H.**
Host Susceptibility Factors Branch
Epidemiology and Genomics Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
National Institutes of Health
Executive Plaza North
6130 Executive Boulevard
Bethesda, MD 20892
(301) 496-8105
mechanil@mail.nih.gov

Jerry Menikoff, M.D., J.D.
Director
Office for Human Research Protections
U.S. Department of Health and Human Services
1101 Wootton Parkway
Rockville, MD 20852
(240) 453-6900
jerry.menikoff@hhs.gov

Stefanie Nelson, Ph.D.
Program Director
Epidemiology and Genomics Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
National Institutes of Health
Executive Plaza North
MSC 7324
6130 Executive Boulevard
Bethesda, MD 20892-7324
(301) 435-6613
stefanie.nelson@nih.gov

P. Pearl O'Rourke, M.D.
Director, Human Research Affairs
Partners HealthCare
116 Huntington Avenue, Suite 1002
Boston, MA 02116
(617) 424-4152
porourke@partners.org

James Ostell, Ph.D.
NIH Distinguished Investigator
Chief, Information Engineering Branch
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Building 45
Bethesda, MD 20892
(301) 435-5978
ostell@ncbi.nlm.nih.gov

Erin Ramos, Ph.D., M.P.H.
Epidemiologist
Office of Population Genomics
National Human Genome Research Institute
National Institutes of Health
MSC 9307
5635 Fishers Lane, Suite 3058
Bethesda, MD 20892
(301) 451-3706
ramoser@mail.nih.gov

***Laura Lyman Rodriguez, Ph.D.**
Director
Office of Policy, Communications and Education
National Human Genome Research Institute
National Institutes of Health
Building 31, Room 4B-09
31 Center Drive
Bethesda, MD 20892
(301) 594-7185
(301) 402-0837 Fax
laura.rodriguez@nih.gov

***Kenna Shaw, Ph.D.**
Director, The Cancer Genome Atlas
National Cancer Institute
National Institutes of Health
Building 31, Room 3A-20
31 Center Drive
Bethesda, MD 20892
(301) 435-3864
shawk@mail.nih.gov

Stephen Thomas Sherry, Ph.D., M.A.
Chief, Reference Collections Section
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
MSC 6510
8600 Rockville Pike
Bethesda, MD 20892-6510
(301) 435-7799
sherry@ncbi.nlm.nih.gov

Anna Smith
Ethical, Legal and Social Implications (ELSI) Manager
Frederick National Laboratory
Cancer Human Biobank
11400 Rockville Pike, Suite 700
Bethesda, MD 20892
(240) 485-6402
anna.smith2@nih.com

David L. Tabb, Ph.D.
Associate Professor
Department of Biomedical Informatics
Department of Biochemistry
Vanderbilt University Medical Center
MRB III 9160
465 21st Avenue, South
Nashville, TN 37232-8575
(615) 936-0380
david.l.tabb@vanderbilt.edu

Sharon F. Terry, M.A.
President and Chief Executive Officer
Genetic Alliance
4301 Connecticut Avenue, NW, Suite 404
Washington, DC 20008
(202) 966-5557, ext. 201
sterry@geneticalliance.org

Jim Vaught, Ph.D.
Deputy Director
Office of Biorepositories and Biospecimen Research
National Cancer Institute
National Institutes of Health
11400 Rockville Pike
Bethesda, MD 20892
(301) 451-7314
vaughtj@mail.nih.gov

Jennifer K. Wagner, J.D., Ph.D.
Division of Translational Medicine and Human Genetics
Center for the Integration of Genetic Healthcare Technologies
Perelman School of Medicine
University of Pennsylvania
1112 Penn Tower
399 South 34th Street
Philadelphia, PA 19104
(919) 619-3884
jennifer.kristin.wagner@gmail.com

Vivian Ota Wang, Ph.D., M.S.
National Human Genome Research Institute
National Institutes of Health
MSC 9305
5635 Fishers Lane, Suite 4076
Bethesda, MD 20892-9305
(301) 496.7531
otawangv@mail.nih.gov

***Carol J. Weil, J.D.**
Senior Advisor for Ethical and Regulatory Affairs
Office of Biorepositories and Biospecimen Research
National Cancer Institute
National Institutes of Health
11400 Rockville Pike, Suite 700
Bethesda, MD 20892
(301) 442-4270
carol.weil@nih.gov

John T. Wilbanks
Senior Fellow in Entrepreneurship
Ewing Marion Kauffman Foundation
1201 Pine Street, Unit 141
Oakland, CA 94607
(510) 735-9909
wilbanks@gmail.com

Barbara Wold, Ph.D.
Director, Center for Cancer Genomics
National Cancer Institute
National Institutes of Health
Building 31, Room 11A-52
31 Center Drive
Bethesda, MD 20892
(301) 496-6613
barbara.wold@nih.gov

Sheila Cohen Zimmet, B.S.N., J.D.
Senior Associate Vice President for
Regulatory Affairs
Georgetown University Medical Center
243 Basic Science Building
3900 Reservoir Road, NW
Washington, DC 20057
(202) 687-8437
zimmets@georgetown.edu

Science Writers:

Melanie Lymon Harris, M.S., J.D.
melanie@roseliassociates.com

Frances McFarland Horne, Ph.D., M.A.
fmhorne@roseliassociates.com

Chandra Keller-Allen, Ed.D., M.P.A.
chandra@roseliassociates.com

Rose Maria Li, Ph.D., M.B.A.
rose@roseliassociates.com

Silvia Paddock, Ph.D.
silvia@roseliassociates.com

APPENDIX 2: THINK TANK AGENDA

Monday, June 11

8:00 a.m. - 8:30 p.m. **Registration**

8:30 a.m. - 8:45 a.m. **Welcome** *Conference Rooms C-F*
Carol J. Weil
Senior Advisor for Ethical and Regulatory Affairs
Office of Biorepositories and Biospecimen Research
National Cancer Institute, NIH

Laura D. Buccini
Associate Professor
Cleveland Clinic and Lerner College of Medicine
Case Western Reserve University

8:45 a.m. - 10:15 a.m. **Keynote Speakers**

8:45 a.m. - 9:20 a.m. ***Is It or Isn't It? Evolving Policy Considerations Regarding Genomic Data and Identifiability***
Laura Lyman Rodriguez
Director
Office of Policy, Communications and Education
National Human Genome Research Institute, NIH

9:20 a.m. - 9:55 a.m. ***Can You See the Real Me? Human Patients and Human Research Participants***
Misha Angrist
Assistant Professor
Duke Institute for Genome Sciences and Policy

9:55 a.m. - 10:15 a.m. ***Questions and Answers***
Moderator: Nicole Lockhart
National Cancer Institute, NIH

10:15 a.m. - 12:40 p.m. **Invited Speakers and Discussion**

10:15 a.m. - 10:40 a.m. ***-Omics and the Changing Face of Identifiability***
Bradley Malin
Director
Health Information Privacy Laboratory
Vanderbilt University

10:40 a.m. - 11:00 a.m. ***Discussion***
Moderator: Stephen Thomas Sherry
Chief
Reference Collections Section
National Center for Biotechnology Information
National Library of Medicine, NIH

11:00 a.m. - 11:10 a.m. ***Break***

11:10 a.m. - 11:35 a.m. ***IRBs: Forced to Deal With"Identifiability" of Everything...a Daunting Mandate!***
P. Pearl O'Rourke
Director
Human Research Affairs
Partners HealthCare

11:35 a.m. - 11:55 a.m. ***Discussion***
Moderator: Sara Chandros Hull
Director
Bioethics Core
National Human Genome Research Institute, NIH

11:55 a.m. - 12:20 p.m. ***Hypothesis-Testing and Hypothesis-Generating Modes of Research***
Leslie Glenn Biesecker
Chief
Genetic Disease Research Branch
National Human Genome Research Institute, NIH

12:20 p.m. - 12:40 p.m. ***Discussion***
Moderator: Erin Ramos
Epidemiologist
Office of Population Genomics
National Human Genome Research Institute, NIH

12:40 p.m. - 1:40 p.m. **Lunch** *Conference Room H*

1:40 p.m. - 2:50 p.m. **Invited Speakers and Discussion** *Conference Rooms C-F*

1:40 p.m. - 2:05 p.m. ***Portable Legal Consent***
John T. Wilbanks
Senior Fellow in Entrepreneurship
Ewing Marion Kauffman Foundation

2:05 p.m. - 2:30 p.m.	***Making Sense of Genomics While Protecting People*** Deborah Collyar President Patient Advocates in Research
2:30 p.m. - 2:50 p.m.	***Discussion*** Moderator: Sharon F. Terry President and Chief Executive Officer Genetic Alliance
2:50 p.m. - 3:15 p.m.	***The Future of Genomic and Health Data*** Kenneth Chahine Senior Vice President and General Manager Ancestry DNA, LLC
3:15 p.m. - 3:35 p.m.	***Discussion*** Moderator: Kimberly Barnholt 23andMe
3:35 p.m. - 3:45 p.m.	***Break***
3:45 p.m. - 4:10 p.m.	***Seeding the Data Commons: Legal Safe Harbors for Research Data*** Jane Bambauer Associate Professor of Law University of Arizona
4:10 p.m. - 4:30 p.m.	***Discussion*** Moderator: Ellen Wright Clayton Craig-Weaver Chair in Pediatrics Professor of Pediatrics Professor of Law Cofounder Center for Biomedical Ethics and Society Vanderbilt University
4:30 p.m. - 5:00 p.m.	***Wrap-up***
5:00 p.m.	***Adjournment***

Tuesday, June 12

8:30 a.m. - 9:30 a.m.	***Summary of Day 1*** **Breakout Group Instructions**	*Conference Room H*
9:30 a.m. - 11:30 a.m.	**Breakout Sessions**	
	Group 1 Conference Room C	
	Group 2 Conference Room D	
	Group 3 Conference Room E	
	Group 4 Conference Room F	
11:30 a.m. - 12:30 p.m.	**Lunch**	*Conference Room H*
12:30 p.m. - 2:00 p.m.	**Presentation of Breakout Recommendations**	
2:00 p.m. - 2:30 p.m.	**Wrap-up** Carol J. Weil Senior Advisor for Ethical and Regulatory Affairs Office of Biorepositories and Biospecimen Research National Cancer Institute, NIH Laura D. Buccini Associate Professor Cleveland Clinic and Lerner College of Medicine Case Western Reserve University	
2:30 p.m.	**Adjournment**	

APPENDIX 3: GROUP 1 DISCUSSION SUMMARY

Questions the Group Considered

What factors should be considered in the development of a Federal policy for access to publicly funded "-omics" research data?

- Are different types and models of access (open vs. controlled vs. hybrid) appropriate for different types and levels of data (individual vs. aggregate, GWAS vs. WES vs. WGS)?
- Is there appropriate justification for treating "-omics" data differently from other types of research data?
- How should Federal policy take into account international data access and privacy standards?
- What research data or analysis is still needed to address these questions?

Issues in Data Access

Data Breaches

Discussions of -omics data privacy and identifiability often focus on data breaches, most often the inappropriate access to and use of information to re-identify an individual. The public is also increasingly suspicious about nonmedical use of their data, such as by law enforcement. However, at present, a more likely scenario involves a legitimate user who downloads data and misplaces or loses the device on which that data are stored, as has happened with the U.S. Department of Veterans Affairs and with several credit card companies. Such scenarios carry the risk for journalistic scrutiny, which for Federal agencies could lead to the public perception that the government does not protect individuals' data. Breakout group participants expressed concern that the protections built into existing Federal databases for biospecimen or -omics data would prove insufficient under such scrutiny.

With data repositories managed by NCI, NHGRI, or other NIH Institutions and Centers (ICs), sponsored research persons (including PIs) sign a statement acknowledging their responsibility in keeping data secure. However, in some cases, investigators then share their access numbers with laboratory staff, or they control and disseminate the password to de-encrypt data. NIH systems allow PIs to designate users or downloaders of data. In any case, the notion that the PI on a project is the only one accessing sensitive data is likely wrong in some cases, and sponsored research persons who sign data use agreements often do not understand the scope of their responsibility to protect patient data.

Limited Access

Although the risk of data breach is concerning to NIH, research institutions, and research participants, there is consensus that legitimate users should not be barred from accessing data needed for research purposes. Students can be blocked from accessing real data, and frustrated by being assigned to work on simulated datasets. Researchers may find it difficult or impossible

to use an important database because of limited access. The Cancer Genome Atlas (TCGA) has developed a database resource separate from NCBI in order to provide different tools to meet users' preferences for manipulating and viewing the data. Some users find dbGaP difficult to use; although they eventually access the data, the additional protections prove cumbersome. In some cases, to facilitate research, stakeholders are working around Federal limitations on data access by developing their own databases. NIH is expecting extramural scientists to share their data, but it is not clear how NIH enforces or facilitates such sharing. Breakout group participants emphasized the importance of balancing risk versus benefit and of not limiting access in a way that severely limits the downstream benefits of data sharing.

The Responsibility of the User

The breakout group agreed that access should be as open as possible but that users should be reminded of their obligation to the persons whose data they are accessing. There should be a more direct relationship between the database and each user, requiring users to identify themselves before they can access the data. Such reminders could come in the form of an additional page or interrupt, an update of the Federal systems that requires individuals at a "licensed" institution to apply for permissions, a formal licensing or roster system with continuing education requirements for individual users, or tiers of access based on broad definitions of who can legitimately work with the data. For example, a database of childhood cancer data would allow tiers of access to researchers working on pediatric cancers and to those who wish to use these data in comparisons.

Data Protections

Suggested data protections include a system in which filters are placed to address issues specific to each type of data. Such a system would require additional education and review of individuals requesting access to more sensitive types of data, and could provide maximum flexibility for researchers by focusing on stringent protections where they are needed most. However, some breakout participants pointed out that "we don't know what we don't know" and that data that do not appear to be sensitive at the time filters are designed could later be found to link to more sensitive information.

Data are particularly more vulnerable during transfer to secondary or tertiary users. Thus safety protocols should be in place as data are transferred to maintain the integrity of the data and allowable uses. This is particularly true in international collaborations, where investigators in different countries might face different privacy standards. Secure cloud storage, where fewer data transfers would occur and computation would be performed within a secure environment, could address risks associated with data transfer and allow systems to monitor data that have been transferred. However, such a model would likely hamper research. A secure cloud would require a large amount of money, bandwidth, and computing power, and it is not clear that a cloud would be able to handle the most critical analyses. Likewise, an app model in which software would create analytic algorithms or provide a process for creating them would be difficult to develop for -omics analyses, which require large blocks of data to be transferred.

Deterrents to Inappropriate Use

Despite everyone's best efforts, not all data breaches can be stopped. Breakout participants noted that even with HIPAA, health information is still exposed and "the CVS down the street is not very secure." By accepting that inevitability, data repositories could focus instead on inappropriate use. Breakout participants cited, for example, pending litigation against two insurance companies who have obtained genetic information legally through the health insurance they underwrite and are now using that information to guide patterns of underwriting for their life insurance business. Use of genomic data by law enforcement was also debated as an example of inappropriate use. Some would argue that law enforcement officials would violate the Fourth Amendment if, upon finding a blood spot at a crime scene, they demanded individuals' genomic data and used it to get to someone through his or her relatives. Others would argue that this use of data would serve interests of public safety and security.

Currently, dbGaP and other NIH databases may penalize inappropriate use by sanctioning access for violators, and potentially if problems are systemic, from the offending institution. However, these mechanisms might not hold individual bad actors accountable. The breakout group therefore agreed that systems should include deterrents significant enough to force individual end users to take responsibility for inappropriate use. Several participants suggested the need for legislation outlawing inappropriate use of genomic data, similar to GINA, but they acknowledged that passage of such legislation is a slow process. Others suggested stipulating in contracts that individual end users would be held liable for inappropriate use. However, others acknowledged that NIH cannot write such contracts because it is a Federal agency and therefore must make its data publicly available and appropriately accessible.

Incorporating deterrents against inappropriate use would require clear definitions of inappropriate use beyond publication of someone's identifiable information. Re-identifying participants from a data archive and selling their information to insurance companies is one example of inappropriate use. Re-identifying and exposing participants from a sensitive study, such as one focused on mental illness, is another. Systems also must define research use clearly. Breakout group participants pointed to the lawsuit by the Havasupai tribe against genetic researchers, where use of genetic information did not necessarily breach the consent document but did breach the tribe's trust, which was based on its understanding of the research.

Need for a Change in Culture

Data access is also hindered by the competitive research culture incentivized by NIH and academic institutions. NIH has built its research databases to allow all scientists prompt access to the greatest extent possible, while complying with study consents and with laws such as HIPAA that are not primarily geared toward research. However, outside of such centralized research efforts with broad data-sharing policies, laboratory investigators may feel uncomfortable revealing data to other scientists. Other investigators might participate in collaborations, but may not share their data until they have published first. In addition, it is not always clear that research participants are truly informed when they consent to have their specimens and information collected. Moreover, for some disciplines focused on phenotypes associated with stigma, such as research on mental illness, addiction, or illicit behaviors, investigators are unwilling to share data. NIH has stated publicly that it encourages data sharing, but the level of sharing varies

across ICs. Other scientific disciplines, such as astronomy, have turned to crowd-sourcing to enable more rapid data sharing, but breakout participants agreed that applying these principles to health research is far more complicated.

The breakout group called for bold leadership to address this challenge. Participants noted the need for a national dialogue to shift the culture in a way that balances societal benefit with individual choices. It also noted current efforts, led by NHGRI, to expect investigators to share certain types of data immediately upon quality control. This model, used in the GWAS policy, allows a time window where other investigators can access the data, but only the original investigator can publish results derived from it. At the end of that buffer period, anyone would be able to publish or otherwise report on his or her results. Other breakout group participants suggested that lessons be learned from pediatric oncology, where investigators have succeeded in sharing data. An inventory of existing cohorts and their levels of data sharing was suggested.

Need for Education

As NIH implements its policies for data sharing, it should engage in community outreach to educate the public about the evolving risks associated with genomic data collected during research participation and the existence of increasing volumes of available data for matching through other publicly available sources. Furthermore, the deterrents in place to address those risks, the problems associated with severely limiting access to data, and the improvements and progress NIH and other agencies have made to overcome those problems need to be communicated. In addition, education is needed to help users understand the responsibilities they take on when they request data. However, it is not clear how best to educate users.

The public should also know about the potential benefit of -omics research and data sharing. Individuals make choices about risk in their daily decision making, and they accept risk in situations where the potential benefit is overwhelming. Breakout participants thus suggested that NIH should communicate with and educate the public, the media, and politicians about the vast potential that could be realized by -omics research and data sharing. At the same time, the group cautioned that NIH should engage in education to raise awareness without overpromising or appearing to be an inappropriate cheerleader for all uses of genomic data. Participants suggested that NIH work with patient advocacy groups, whose stance and loyalty to the patient is beyond reasonable doubt, to increase public awareness.

Need for Resources

Because of the way the NIH genomic data-sharing databases were built, some group members felt that they are not scalable and are therefore ill suited for the recent dramatic growth in the amount of data, the number of users, and the diversity of uses for the data. The amount of data coming in is starting to overwhelm these systems, creating a bottleneck to access. Breakout group participants thus called for additional full-time employees, infrastructure, and financial resources to manage the data and for bold leadership to navigate roadblocks. However, commitment of resources should be practical and done with realistic expectations. One breakout group participant summarized this sentiment by suggesting that NIH "tweak, but not redo."

A Granular System of Privacy Controls

Nationwide consistency on privacy controls will likely never be reached; a granular system of such controls is therefore needed. The Genetic Alliance, in collaboration with Pfizer and others, has built such a system where research participants can check what they are and are not willing to share and whether they consent to be re-contacted. Guides with models of varying comfort levels have been developed to guide patients in setting controls, and the list of scenarios in the system is extensive. Pilot projects have been carried out for patients with psoriasis and Kleinfelter syndrome, with about 90 percent compliance for each group. However, Pfizer has dissolved its innovations department and therefore no longer supports this system, and the Genetic Alliance has not yet been able to find new support. Many potential funders have expressed excitement over such a system but remain limited by constrained budgets. Efforts to secure funding are ongoing.

Proposal: A Pilot of an Open-Access Model

As demonstrated by several presentations at the think tank, some patients and other research participants are willing to contribute and share their data if the benefit of sharing outweighs the risk. However, although breakout group participants agreed on the need for policies, infrastructure, and incentives to support those willing to share their data openly, they acknowledged that comfort levels vary across different populations. The breakout group therefore suggested a demonstration or pilot project to build a longitudinal cohort of research participants willing to donate their specimens and data to an open-source data repository. The consent document for this cohort would state clearly that data collected during the study would be made public, that anyone would be able to access the data, and that there was a risk, although small, that participants and their conditions could be re-identified. Such a project would be intended not to replace the existing research model, but to see if a truly open-access model would work. The breakout group hoped that such a project would demonstrate the widespread benefits as well as the acceptability of open access and data-sharing and thus guide educational efforts about the potential of -omics research and sharing. Group members also accepted the possibility for stratified recruitment, based on comfort level, and even suggested a series of pilot projects or a stratified gateway model to address varying patient cohorts or privacy levels. However, the group emphasized that if resources are limited, one pilot project on an open-access cohort would be a good start.

The International Landscape

The breakout group briefly discussed differences in privacy standards across countries. Although members agreed that data repositories should not blatantly violate these standards, they did not feel that privacy and identifiability policies in the United States should be tailored to fit them.

APPENDIX 4: GROUP 2 DISCUSSION SUMMARY

Questions the Group Considered

What considerations enter into determining whether -omics data are identifiable?

- What distinguishes "identifiable" data from "de-identified" data?
- Can data ever truly be "de-identified," or is that concept outdated in the genomics era?
- What criteria or standards should be used to establish whether particular types of -omics research produce identifiable or de-identified data?
- What research data or analysis are still needed to address these questions?

Barriers Restrict Scientific Discovery

There was an overall consensus that some barriers concerning identifiability are necessary to prevent abuse. However, the more barriers are put in place, the more scientists are restricted from moving science forward, resulting in less discovery. Presently, "-omics" data are not legally identifiable, but they certainly can be used to identify participants. Therefore, NIH has sought to protect those types of data that clearly are identifiable, such as whole genome sequences, and to provide protection from free and open distribution of data. Although these protections are necessary, they raise concerns about shutting down the free exchange of information.

All Data Are Identifiable

Conflicts exist because researchers submit data to gene expression databases, even as recent papers report that expression patterns can be identifiable, provided that a matched DNA sample is available. For the purposes of discussion, the breakout group thus assumed that the majority, if not all, data are identifiable, unless protections are put in place to make them unidentifiable. However, as time progresses, identifiability requires a link back to other datasets, such as vital signs, ribonucleic acid (RNA), or other genomic products; the level of identifiability depends upon the amount of data available, and some types of data are at greater risk than others for identification. Moreover, the context in which data are used, and in particular who uses the data, must be considered. Broader policy issues arise from the potential identifiability of data and the potential ability to connect someone's -omics data to other types of data.

A credit card analogy was used to suggest that liability for identifiability be capped for patients or other research participants but shifted heavily toward researchers and institutions. Such a shift would maximize access to data. The breakout group also discussed the criminalization of unpermitted reverse identification from genomic data and suggested that the breach of data be treated as the equivalent of medical malpractice for researchers. However, how to enforce such a policy is not clear. Any punitive policy should exempt re-identification that is pre-approved and conducted for the benefit of the research participant.

Russian Doll Model of Access

Some data repositories, such as NCBI, use a model of three tiered levels of consent—open, generally controlled, or restricted access—that are not based on identifiability. Open access, which has no restrictions and is available for nonhuman data and gene expression, provides a mechanism, other than the IRB consent process, to use data that come from outside of NIH. Thus open access maximizes the number of users. Generally controlled access, which is used for data from nonhuman subjects research, involves application to a central resource with consent for broad sharing. Restricted access applies to data from at-risk populations who have not consented for their data to be available for broad use.

The breakout group suggested streamlining the generally controlled access level by establishing one DAC for general research use. Such a process would employ a general consent for research use, thereby eliminating the need for new studies to undergo consent and increasing access to data. A trusted partner, which is better able to serve targeted communities, also could facilitate the sharing of restricted or generally controlled data. However, the trusted partner process places archival longevity at risk, is poorly integrated with other outside NIH resources, and is expensive and complicated. Moreover, because of the complexity of the trusted partner process, a partner can exist outside NIH but still be under NIH control/legal obligation.

Appropriate Consent

The assumption that all data are identifiable and the concerns of patients lead to the requirement of informed consent for all research uses of data. However, it is important to consent participants in an appropriate way. Appropriate consent requires not only the simplification of the consent document, but also the removal of arbitrary restrictions. For example, the Australian model delegates consent to the IRB such that the IRB decides whether a proposed study involves the ethical use of a person's sample. The delegation of consent to a group broadens patients' ability to authorize use of their samples in ways that they cannot predict at present (including the ability to link samples to participants), and it gives patients the right to opt out.

Best practices and standardized consent documents are needed for data sharing and informed consent with respect to samples, data, and Internet use. Breakout group participants therefore suggested that a standard consent template be added to the *NCI Best Practices* document and that, if necessary, a non-mandated standardized data-sharing plan be created and linked with that template. Although the requirements for informed consent at each access level need to be determined, the breakout group suggested that templates include minimum restrictions such that standardized data-sharing plans could be customized. The group also suggested that a community advisory board be established to address issues of consent and to explain the risks and benefits of research studies, as well as the ability of community members to veto or opt out of the study. Such a board could build public trust through community engagement.

Law, Policy, Procedure, Practice

The breakout group discussed the separation of policy from procedure and the relationship between the way a system works and how one realizes and facilitates access. IRBs need guidance on ways HIPAA and the Common Rule intersect, but they are also influenced by anticipated

changes to current policy. In addition, with the assumption that all data are identifiable, and in order to protect identifiable data while allowing researchers access, it is important that the policies and regulations in place make it difficult to violate research participants' privacy and reveal their identity. By getting involved in policymaking and implementation, scientists could better understand the connection between law, policy, procedure, and practice.

Suggested Next Steps

On the basis of its discussion, the breakout group suggested several next steps:

- More research data and analysis are necessary to determine whether -omics data are identifiable. It is not clear whether NIH should support such research or wait for the problem of identifiability to arise.
- Streamlined access should be provided to the generally controlled data, for example by establishing fewer and better educated DACs.
- There is a need for broader community engagement that provides decision-making power to community members and gets scientists more involved in policymaking.
- A broader discussion regarding potential identifiability needs to be held with the proteomics community.
- Proteomic data should be treated as unidentifiable, and thus included in the open-access tier of the Russian doll model, until evidence shows them to be otherwise. A task force should be established to look at ways to address the production of identifiable data from proteomics research.

APPENDIX 5: GROUP 3 DISCUSSION SUMMARY

Questions the Group Considered

What are the appropriate ethical constraints to allowing researchers broad access to -omics data?

- What do we know about participant attitudes toward investigator access to their DNA and the privacy-utility tradeoff of limiting data access?
- What do research participants and the public actually understand about the use of DNA in research (e.g., growth of cell lines, induced pluripotent stem cells), and what should they be informed about before consenting to participate?
- To what extent should the concepts of autonomy, beneficence, and justice limit access by researchers to an individual's -omics data?
- What research data or analysis is still needed to address these questions?

The Altruism of Research Participants

Many patients, and even healthy family members, will place the science ahead of their privacy concerns if they are informed, feel involved in the research process, and trust the researchers. However, attitudes and levels of acceptance vary widely among research participants. For example, healthy participants will have different attitudes than those with diagnosed diseases and conditions. Empirical evidence indicates that participants prefer to be asked up front for their participation in research and tend to be upset if they find out after the fact that their specimens are used for research. They want their altruism to be valued, and they want to know how their donations are used. Breakout group participants thus focused on the importance of building trust between researchers and research participants, although some felt building trust is impossible because there is no oversight or control over what researchers do with specimens or data after they obtain them. The group disagreed, for example, on whether a situation similar to what happened to Henrietta Lacks[62] could happen in today's research atmosphere. Some group members felt strongly that the current rules, if followed, would preclude such a scenario, whereas others felt equally strongly that such a scenario could still happen today.

Once patients are informed about the types of research done with biospecimens, they likely will want access to validated research results. However, the current research enterprise does not enable such access on a broad scale. An option for sustained interactions between researchers and participants should be incorporated.

Distinctions between Basic and Applied Science

The group discussed the distinction between basic science, where specimens or -omics data are typically not identifiable because names or personal health information are not linked to them, and applied research, where it makes more sense to maintain identifiability because of possible

[62] http://www.archivesofpathology.org/doi/full/10.1043/1543-2165-133.9.1463.

therapeutic implications. How to handle the perceived distinction was a subject of continued debate. Some group members felt strongly that if all tissue samples are considered identifiable in a future system, basic science applications would suffer because of the added onus of obtaining consent. They suggested that discardable specimens (i.e., surgical waste) used for basic science research with no therapeutic implications should be protected from any re-identification, because there would never be a reason to return results to patients or contact them for more information or for return of results. Other members argued that all specimens, whether labeled with explicit identifiers or not, are inherently identifiable. These members thus suggested that obtaining consent from all research participants addresses an imperative to show respect for all persons, regardless of whether their specimens and data are used for basic or applied research, and regardless of whether informed consent is required by current or future regulations. The balance between protecting privacy and promoting utility, or the question of how the research enterprise could incorporate basic scientific discovery work while acknowledging the possibility that samples could someday be re-identified for as-yet unknown innovations, was not resolved.

Maintaining Identifiability and an Appropriate Privacy-Utility Balance

Many group members did agree that the future of research lies in the benefits gained from maintaining identifiability of all data and thus enabling researchers to return research results and seek additional information that could inform further research. They added that basic research applications that appear to have no therapeutic implications and thus no need for identifiability could have such implications in the future. At a minimum, hypothesis-generating and hypothesis-driven research should coexist in the research environment, and a mechanism should be in place to obtain and track consent for research using specimens and data.

At present, tissue specimens are often processed through a biobank, which provides researchers with de-identified data and/or specimens but maintains a coded key to re-identify samples. Although this mechanism protects the research participant while allowing for the possibility of re-identification to obtain information or return results, the tracking systems in place at biobanks might be insufficient to guarantee sample identifiability, and the chain of custody of specimens may be inadequate. Better accountability for maintaining identifiability at biobanks will be needed to accommodate a consensus that all samples are identifiable.

It is legally safer for researchers to work with de-identified data. But the liability of the researcher is unclear in cases where research results are ambiguous or incorrect, and an assumption that all data or biospecimens are identifiable could place undue responsibility on the researcher. The research community exacerbates this ambiguity, providing a strong disincentive for using identifiable data. However, researchers who do not yet embrace the hypothesis-free research design presented by Dr. Biesecker might not fully understand the magnitude of loss of capabilities that results from using de-identified data. Thus clinicians and researchers need to be educated about the tradeoffs between privacy and utility.

Providing Potential Research Participants with Information

The breakout group agreed that in general, the public is not scientifically literate. Patients do not often see or understand the distinction between research and treatment. They are asked to sign a variety of consent forms for several reasons, but they do not always read them carefully,

understand them, or distinguish between the types of consent. However, scientific and research literacy even among clinicians is of some concern, particularly with respect to stem cell research. Initiatives to educate the general public about research outside of the clinical context are needed, in addition to processes where information is shared with patients or research participants at the time they sign consent. However, as with smoking, high blood pressure, or obesity, education and information will not necessarily change behavior.

The consent process should communicate clearly and transparently the current knowledge base, the research gaps, and the role of research participants' donations in addressing those gaps. The process should help participants to understand that they are donating specimens or data for research that may or may not be used to identify genes. Whether or not participants should have the option to give explicit consent for genetic versus non-genetic research was a point of disagreement during the discussions.

Tracking Consent for Downstream Research

The current system does not ensure that consent configurations are tracked and adhered to when specimens and data collected for one study are used by other researchers downstream. In addition, the research enterprise has done a good job of protecting the identity of research participants, but it has been in a defensive crouch, and the intense concerns raised during the 1980s and 1990s about genetic privacy have not borne out. Some breakout group members suggested that secondary researchers should have to re-consent participants and that specimens and data should thus be identifiable. Others suggested that research should be a transparent, participant-driven process that shifts away from a focus on privacy and de-identified data to one where participants are informed of protections and have the responsibility of indicating their preferences. In this environment, the consent process would describe clearly the procedures in place to protect participants' information, but it would not guarantee complete confidentiality, and participants would have access to information about who has accessed their data and how their data are used. 23andMe and the PGP[63] use such approaches, and 90% of 23andMe participants have consented to have their data used for genetic research. However, it was not clear to the breakout group whether a broad consent process would be allowable under the ANPRM.

Guiding Researchers' Access to an Individual's -Omics Data

Breakout group members noted that the question about the extent to which the concepts of autonomy, beneficence, and justice should limit access by researchers to an individual's -omics data assumes that access to data should be limited and that access for legitimate reasons could be unnecessarily limited by protections against the possibility of harm. Autonomy provides rules that govern access based on individual preferences. Thus group members suggested reframing the question to consider what should "govern" or "guide" researchers' access to data.

Respect for autonomy should be operationalized by allowing research participants to make informed decisions and engage in a meaningful thought process about the type of consent they are giving. The standard surgical consent, which in some hospitals only mentions the possibility

[63] http://www.personalgenomes.org/.

that surgical waste will be used for research, constitutes notification, rather than informed consent, and patients, knowing they must sign the surgical consent form to receive treatment, often have no opportunity to deny consent for tissue collection for research. Respect for autonomy also should enable research participants to control as much of the consent process as possible. If a national health information infrastructure that is based on electronic health records and holds patients' information is developed, patients should be responsible for setting the terms and limits governing the use of their information for research. However, to increase patients' trust, the infrastructure will need a tracking system that respects their preferences and allows those preferences and details regarding their consent to follow their sample.

Justice ultimately means maximizing the utility of research for everyone, including underrepresented groups, and maximizing research that will lead to good outcomes for people across a range of diseases. There was some disagreement about the implications of a transparent, participant-driven system on the ethical principle of justice. Some group members suggested that this type of system would limit participation by a segment of the population that does not have easy access to computers, is not computer savvy, does not have time to participate actively, or lacks the resources to understand its role in the research process. These limitations could reduce the representativeness of biospecimens and data. Other members suggested that giving individuals control over their preferences online could actually increase representativeness by allowing people to participate even if they live in rural or remote areas without easy access to medical facilities. Yet other members noted that the system does not have to be completely participant-driven online or completely researcher-driven in the traditional sense. Instead, the research community could prioritize a shift toward participant-driven procedures while maintaining opportunities for alternative procedures that are currently used and might do a better job of reaching disenfranchised populations.

More Research Data or Analyses Needed

More data and analyses are needed to determine how representative the biospecimen research population is. Improved tracking systems that identify which participants agree to what types of research could enable analyses of the populations to determine whether data are skewed. Likewise, longitudinal data about research participants would enable analysis of how successful recruitment is across broad populations. Other studies could engage research participants to learn what information they want to receive from the research studies they participate in and to explicitly describe the tradeoffs of participating. Participants could be educated about resource constraints and the costs associated with robust return of results. After having acquired the necessary knowledge, participants can make educated choices regarding the tradeoff between, for example, conducting two studies with intensive return of results or five studies with less robust return of results. Such an exercise could provide insight into whether participants trust researchers and see the value in more or less research. In another example, a study of members of the "Patients Like Me" community,[64] which has no explicit guarantees of privacy, could ask whether participants have regrets, whether they perceive identifiability positively, and whether the community is representative of patients in general.

[64] http://www.patientslikeme.com/.

Research using biospecimens and -omics data also could benefit from process management case studies that illustrate how the system operates currently and how it would operate under the proposed new rules. These case studies would highlight problem areas in the newly proposed process and enable the development of focused solutions.

Additional Concerns

The breakout group generally agreed on the importance of a system that accommodates both future innovations and capabilities and research uses that have not yet been identified. Group members also agreed that all specimens and data are or will be inherently identifiable. However, some members suggested making changes to the current overall research system, whereas others considered designing an entirely new system that accounts for unknown future capabilities and uses for genetic research. Either way, changes to the research infrastructure would require additional resources, presenting a challenge in light of the current austerity with respect to research budgets. In addition, how to treat existing specimens collected during past studies remains a challenge.

The breakout group also noted that clinical care is strongly governed, with a system in place to address breaches of ethics. However, the research system lacks the same type of oversight. Researcher abuse is not as well governed, and there may not be adequate disincentives or consequences for those who do not adhere to ethics guidelines. Yet a well-publicized case of researcher abuse is a political risk and would likely result in additional regulations and constraints that hamper biomedical research overall.

Data-sharing is needed to maximize the benefit of research using biospecimens or -omics data. If multiple researchers use the same tissue specimens from the same biobank for a variety of research purposes, then the results should be combined in some way to create a larger picture of knowledge. Knowledge that is combined across researchers and projects will have a greater impact than the current fragmented system of research.

APPENDIX 6: GROUP 4 DISCUSSION SUMMARY

Questions the Group Considered

How can society minimize any risks and maximize any participant benefits of -omics research?

- What are the risks of various -omics research technologies and the data they can produce?
- How can -omics studies be designed to maximize individual participant, family, and community benefits (e.g., the return of individual or group population research results)?
- What public or regulatory policies would promote appropriate balance of the risks and benefits of -omics research and help to avoid unwanted disclosures of identity and future uses of DNA for undesired purposes?
- What research data or analysis is still needed to address these questions?

The Risks of Various "-Omics" Research Technologies and the Data They Can Produce

Historical Context

-Omics research is still a very young field. Before this technology was available, research was carried out for many decades in the context of genetic epidemiological studies. PIs collected samples usually from a local catchment area; often they developed long-lasting, personal relationships within this community. The data always stayed with those same investigators, who were usually very selective regarding collaboration. Active efforts to keep the community and relationship with the researchers closed were common.

In sharp contrast, data are now obtained and released at unprecedented speed. Most studies are carried out as large, tax-money funded, collaborative endeavors, and many members of the research community get access to the result and benefit from the initial investment in data analysis. The PI usually has a right to publish first, during a proprietary period, after which the entire dataset gets released to a large number of other, qualified scientists.

Breakout group participants noted that a lot of learning will be necessary in order to develop a culture of shared resources. Additionally, researchers need to consider that a consent form filled out in the 70s or 80s that stated that "data may be shared with other investigators" was likely signed without any concept of today's technological possibilities. Naturally, the increased ability to generate and share data generates a multitude of new issues regarding identifiability and privacy.

Risk and Harm

Breakout group participants found it difficult to restrict the use of the term "risk" in this context to "the risk for being identified." Other important risks included the risk for investigators to violate policies they may not know about, and the risk that data might reveal additional information about an individual who has already been identified. Furthermore, participants noted

the risk that family members of someone who carries a disease-related mutation might learn about his or her carrier status when no treatment is available. Conversely, researchers might face the risk of learning about a patient's disease status without being able to communicate that status to the patient.

The discussion about risk triggered the question of who might be harmed. The breakout group agreed that the actual risk that a person might be harmed by identification based on genomics is very low. The harm, however, might occur in the form of loss of public trust, and the subsequent loss of support for research might exceed the harm suffered by the identified individual.

There was general agreement that all -omics data could be used to identify a research participant as long as a matching sample (derived, for example, from a discarded cup) was available. Some -omics data, such as microbiomics, could reveal additional information about a participant, such as possible exposure to certain infectious agents, and some types of -omics data (e.g., genomics) are more stable over time than others (e.g., metabolomics).

The participants discussed the need to distinguish between probabilistic predictions (e.g., an individual belongs to a certain group or has a higher risk for certain diseases) and absolute ones (e.g., individual is John Doe and has been identified to be a participant in a schizophrenia research study). Individual research participants who enroll in multiple studies are regularly identified as duplicates and excluded from analysis without being identified at the level of personal data. At the family level, for example in the diagnosis of Lynch syndrome, the research result may be highly predictive of disease not only in the carrier, but also in additional family members who might benefit from genetic testing and preventive health measures. However, it is not possible to contact these family members directly without going through the initially tested research participant. It was noted that it was generally believed that these participants understood the importance of notifying additional relatives.

In any case, breakout group participants felt that data security is likely to be more important for protecting an individual's privacy than for protecting against identifiability.

Inappropriate Use

The participants acknowledged that there appeared to be a substantial fear in the population that "something bad may happen" if data fall into the wrong hands. Hypothetical misuses range from elimination of political opponents and other attempts to stigmatize certain people or populations to exclusion from health insurance. However, the breakout group also noted that a lot of epidemiological data are being collected and analyzed and that these datasets could pose as much of a problem for identifiability and privacy as -omics data do. Yet, there seem to be special yet hard to justify security concerns regarding genetic data.

Cases of actual inappropriate use have, indeed, occurred. In the well-known case of the Havasupai tribe, DNA collected to study a certain disease was later used for ancestry and inbreeding studies, which had not been authorized by the tribe members, and the university at which the research had been conducted was held responsible and fined. Public trust may be endangered when conditions specified in the informed consent procedure are neglected in this manner. In addition, anonymization of the dataset is futile in this case, because the population frequencies of the alleles will always make it possible to identify it as this population. Thus this

case illustrates a need for greater awareness of the rules. In another example, the media has reported several cases in which people used -omics data to test for paternity or adoption status without the consent of the biological or adoptive parents. However, because there appeared to be no harm in at least one of these cases, the breakout group did not agree on when to classify these cases as misuse. Breakout group participants also recalled several cases of genetic discrimination and noted that current legislation is sufficient to punish the perpetrator. However, these participants also speculated that a large number of unreported cases might never go to trial, and they pointed out that GINA no longer applies after an individual has developed symptoms of a disease.

How "-Omics" Studies Can Be Designed to Maximize Individual Participant, Family, and Community Benefits

The Consent Process

Because different individuals have different thresholds for what they perceive to be adequate uses of their data, dbGaP has already implemented a tiered access system in which some data are available only for disease-specific studies while others are available for general research use. In the future, even broader access categories may allow researchers to carry out methodological studies in large pooled datasets of many individual studies. Breakout group participants agreed that it is the choice of the individual in which dataset to participate. However, they also noted that once these data are released, withdrawing consent is no longer possible. Furthermore, some research participants might not object to any specific kind of scientific research question but they might exclude their samples from commercial use. Thus the breakout group emphasized the need for consent documents to inform research participants of the possibility that their data might be used for commercial development.

Reporting Research Findings to Other Researchers and the Public

Several breakout group participants stressed the general importance of releasing aggregate or group-level results in public journals and databases. However, they noted that in the future, such releases need to make a better effort to acknowledge all participating clinicians, researchers, and other staff, so that the final team analyzing the data does not take all the credit. The breakout group also suggested that public trust could be built by issuing newsletters regularly and by releasing "good news" resulting from -omics research before negative reports dominate the news media. Breakout group participants noted that the media often has greater influence than researchers on public opinion and trust. They suggested that positive publications illustrating the new abilities to fight common diseases should be released in general. If good news coming from -omics research is released only in response to scandals, the negative image raised by the scandal might be difficult to overcome. Some breakout group participants also pointed out that a policy of reporting findings, for example the discovery of new disease-related alleles, in general might actually reduce the stigma associated with specific variants.

Return of Results to Individuals

The breakout group decided that an exhaustive analysis of the "return of results" issue was beyond the scope of the current discussions. However, the group briefly discussed the consequences of returning results at the individual and group levels. Although the breakout group agreed that how much data is expected back from a study varies widely across individuals, well-informed research participants might benefit from results and demand to be "in control" of their data and health. However, even seemingly simple tests can produce errors or results that are difficult to interpret. No standards have been developed to determine how to deal with errors or difficult-to-interpret results, when a result is considered predictive, and when action should be taken.

Public or Regulatory Policies to Promote Appropriate Balance of the Risks and Benefits of "-Omics" Research

The breakout group agreed that holding users of -omics data accountable will be easier if they have to submit their names and institutional affiliations when downloading data. The current dbGaP controlled access model was cited as one example of holding institutions accountable for the research carried out by their employees. The breakout group also suggested the creation of DACs specifically to manage access to more sensitive data (e.g., drug abuse) and thus relieve existing disease-specific DACs of the burdens and responsibilities associated with these data. Several universities and centers go beyond what is required by the law and have, for example, implemented additional opportunities for individuals to opt out of having their data used for certain types of studies.

In the future, access to several datasets all over the world through one centralized portal would enable researchers to carry out studies in combined populations of unprecedented size. However, this type of portal would require an international harmonization of access rules, and group members noted confusion over how the terms "public," "limited," "restricted," and "controlled" are used.

In order to train bioinformaticians and be prepared for the next great challenges in data analysis, it is essential for trainees to practice on real-world, large-scale datasets. However, obtaining access to such data for training purposes is currently very cumbersome. Availability of dedicated, broadly consented datasets just for training purposes would be desirable. Breakout group participants also noted that obtaining consent for all de-identified samples to be used in research could be prohibitively expensive for assay developers. However, they pointed out that pharmaceutical companies should have the same obligations to deposit research results into public repositories if they are allowed to access to samples and data under the same rules as academic researchers.

More Research Data or Analyses Needed

Education

The breakout group suggested that standards be established to describe to research participants the risks and possible consequences of inappropriate use. The use of a common language in

consent forms for risk and consequences could help in overcoming inter-center variability. Likewise, researchers need to increase their awareness of identifiability and privacy issues, which increase with larger and shared datasets. Individual files may be considered de-identified, but a combination of these may constitute a breach of privacy laws. In addition, to ensure public trust, researchers need to be aware of all the negative consequences, both personal and public, of inappropriate use of -omics data.

Clear Rules and Consequences

For researchers to increase their awareness of privacy and identifiability issues, rules must be more transparent regarding punishments that go beyond lawsuits and loss of reputation. In addition, because it will never be possible to prohibit data sharing or to ensure that all access is monitored, breakout group participants suggested that criminalizing misuse might be important to avoid unwanted disclosures. However, it is not clear whether this approach would deter hackers from unauthorized access to restricted data. Breakout group participants also noted that any law applicable in the United States might be circumvented by individuals in other countries who might access data illegally. They suggested, therefore, that legislation would also need to criminalize downstream uses and preclude global misuse of data to the extent possible.

To avoid a constant need to revise the rules, it is important to keep in mind that what is difficult today technically will be easy in a few years. Researchers and research participants should also keep in mind that once information has been released, it is essentially impossible to retract.

Understanding Individuals' Attitudes Toward the Risks and Benefits

Openness toward data sharing varies widely across research participants. Some patients might actually consider releasing their data a means of having control over data-sharing decisions. However, there is no test that can measure which participants would accept which levels of risk. The ones that participate most often in debates about data sharing are usually highly self-selected, whereas those who prefer to remain anonymous do not participate in these discussions. In addition, some of this inter-individual variability might be explained by age. Breakout group members noted that younger individuals might be more agreeable to higher risks. On the other hand, older individuals might feel that the benefits of their data for the health of coming generations outweigh any risks that they themselves might encounter. The breakout group suggested that a new informed consent process be developed to account for these differences.

Breakout group participants noted that some individuals in each consent group in the Framingham study objected to their samples being used for commercial research. However, this group or participants changed over time. Although the breakout group suspected that issues raised by the media presently play a considerable and likely too prominent role in these considerations, more research is needed to better understand what drives different research participants and communities to exclude their samples from certain types of research.

NIH…Turning Discovery Into Health®