NCI think tank concerning the identifiability of biospecimens and “omic” data

Carol J. Weil, JD¹, Leah E. Mechanic, PhD, MPH², Tiffany Green, MHS, MPH², Christopher Kinsinger, PhD³, Nicole C. Lockhart, PhD⁴, Stefanie A. Nelson, PhD², Laura L. Rodriguez, PhD⁵ and Laura D. Buccini, DrPH, MPH^{2,6}

Purpose: On 11 and 12 June 2012, the National Cancer Institute hosted a think tank concerning the identifiability of biospecimens and “omic” data in order to explore challenges surrounding this complex and multifaceted topic.

Methods: The think tank brought together 46 leaders from several fields, including cancer genomics, bioinformatics, human subject protection, patient advocacy, and commercial genetics.

Results: The first day involved presentations regarding the state of the science of reidentification; current and proposed regulatory frameworks for assessing identifiability; developments in law, industry, and biotechnology; and the expectations of patients and research participants. The second day was spent by think tank participants in

small breakout groups designed to address specific subtopics under the umbrella issue of identifiability, including considerations for the development of best practices for data sharing and consent, and targeted opportunities for further empirical research.

Conclusion: We describe the outcomes of this 2-day meeting, including two complementary themes that emerged from moderated discussions following the presentations on day 1, and ideas presented for further empirical research to discern the preferences and concerns of research participants about data sharing and individual identifiability.

Genet Med advance online publication 11 April 2013

Key Words: biospecimens; data sharing; identifiability; research participant engagement

Prompted by mounting concerns about the advancing science of reidentification, the National Cancer Institute (NCI) convened a group of experts to explore the scientific, ethical, and human participant dimensions of the identifiability of biospecimens and multiple types of “omic” data. The “omic” research fields are stimulating discovery and progress in medicine by leveraging highly annotated data sets in the generation and evaluation of numerous hypotheses with impact across a wide range of phenotypes. To foster the most productive and efficient use of research data to generate and test new hypotheses not necessarily envisioned at the time of data collection, the NCI endorses widespread sharing of data across different areas of research. This support for broad data sharing is exemplified by the National Institutes of Health (NIH) data-sharing policy for genome-wide association studies and anticipated policy for genomic data sharing, which seek to facilitate broad access by researchers to deidentified data in publicly supported or conducted studies, consistent with research participant consent.^{1,2}

Provocative publications, however, have questioned the conventional wisdom that individuals included in data sets of aggregated molecular information are fully anonymous. These studies indicate that DNA variant data from pooled single-nucleotide polymorphism disease studies,³ data sets of

RNA expression levels in tissue samples,⁴ or other quantitative traits⁵ can be linked directly to human research participants if a matched sample is available. Moreover, some investigators suggest that human beings can be uniquely identified from just 30–80 statistically independent single-nucleotide polymorphisms.⁶ Furthermore, a recent study demonstrated the ability to reidentify individuals without the availability of a matching sample. Researchers identified individual research participants using publicly available whole-genome research data to calculate short tandem repeats on the Y chromosome in order to query publicly accessible genealogy databases and infer surnames. The genomic research data were linked to metadata, which enabled the investigators to deduce personal identities.⁷

Evolving recognition of the increasing challenges for deidentification of human tissue and associated data inspired a federal regulatory proposal by the US Department of Health and Human Services (DHHS) to consider categorizing all biomedical research involving biospecimens, including collection, storage, and secondary analysis of existing tissue, as research involving identifiable data.⁸ The DHHS sought public comment on potential advantages and disadvantages of considering biospecimens and associated data inherently identifiable. The DHHS also solicited response as to whether specific data

¹Diagnostics Evaluation Branch, Cancer Diagnosis Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA; ²Host Susceptibility Factors Branch, Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA; ³Office of Cancer Clinical Proteomics Research, Center for Strategic and Scientific Initiatives, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA; ⁴Division of Genomics and Society, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA; ⁵Division of Policy, Communications and Education, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA; ⁶Digestive Disease Institute, Cleveland Clinic, Cleveland, Ohio, USA. Correspondence: Carol J. Weil (carol.weil@nih.gov)

Submitted 20 September 2012; accepted 25 February 2013; advance online publication 11 April 2013. doi:[10.1038/gim.2013.40](https://doi.org/10.1038/gim.2013.40)

security protections, including encryption and periodic retrospective random audits, should extend to research using certain types of genomic data, such as genome-wide single-nucleotide polymorphism analyses or whole-genome sequences.

Ultimately, the ability to protect research participants from any unauthorized identification of their biospecimens or “omic” data is an issue of public trust that the research community must address forthrightly and transparently. The NCI must be proactive in contemplating the extent to which evolving technologies might undermine the protection of data that is currently deemed “deidentified” under the law and regulations. We wished to consider whether the burgeoning volume of “omic” data generated in biomedical research, coupled with evolving linkage capabilities posing potential informational risks to research participants and genetically linked populations, warranted a reexamination of informed consent disclosures or data-sharing policies. This article reflects a summary of the discussion from the NCI meeting and is not intended to represent an opinion of the NCI, the National Human Genome Research Institute, the NIH, or the DHHS.

SETTING

On 11–12 June 2012, we gathered a group of leaders in the research community with varied relevant backgrounds (**Supplementary Appendix S1** online) in order to convene a think tank on the science and policy of reidentifying human biospecimens and “omic” data. Our aim was to attempt to outline considerations for the protection of “omic” data and for the distribution to researchers of stored biospecimens and associated data within appropriate ethical and regulatory parameters. The agenda involved an initial day of presentations by leaders in medical research, bioinformatics, ethical and regulatory policy, consumer genetic testing, and patient advocacy. The presentations discussed current challenges posed by biotechnological advances and their effect on participant privacy and perceptions about data sharing. Moderator-led discussion followed each presentation, enabling think tank participants to share experiences and identify challenges and opportunities for empirical research. On the second day, participants convened in one of four preassigned breakout sessions (**Supplementary Appendix S2** online) composed of individuals with varying expertise, for in-depth discussion of specified subtopics.

MAJOR FINDINGS

State of the science

The purpose of day 1 of the think tank was to report on the state of the science from a variety of expert perspectives and, specifically, to address current challenges surrounding the identifiability of “omic” data and biospecimens. There were a total of nine state-of-the-science presentations, including two keynotes. These talks provided background in multiple fields of expertise, including cancer genomics and proteomics, bioinformatics and statistics, legal and regulatory guidance, and industry applications of gene technologies. A brief description and highlights of the presentations are provided (**Table 1**).

Breakout sessions

Day 2 was devoted to breakout sessions with groups focused on specific questions posed by the think tank organizers (**Supplementary Appendix S2**). These deliberations framed suggestions for next steps, designed to move the science forward, while helping to address policy concerns and develop best practices around identifiability. Outcomes of the breakout group sessions are highlighted in **Table 2**.

Two broad themes emerged from think tank discussions about the identifiability of human tissue and data. The first theme concerned the issue of whether additional institutional or legal measures were necessary to address either negligent or intentional violations of research participant privacy. The second theme concerned the extent to which greater understanding and expansion of data-sharing options might provide an opportunity to increase the engagement of participants in the research enterprise.

With regard to the first theme, some experts argued that current governance strategies such as honest broker coding models and irreversible deidentification systems, although not immune to mishap, function well in many institutional biorepositories. According to this view, although it has indeed been demonstrated that with targeted effort, a researcher may infer specific information about an individual from aggregated data (such as cohort participation in a case or control group), this should not condemn current deidentification approaches and best practices. Although often highly publicized, incidents of reidentification from trace amounts of individual data within aggregated data sets are usually isolated and thus need not warrant rewriting of research consent disclosures about privacy, or constriction of researcher access to data. Although misuse of genetic information by researchers is demonstrably feasible, it constitutes such an extreme breach of medical and professional ethics as to best be handled by targeted professional measures rather than by broad policy strokes. For example, wrongdoers might potentially be banned from participation in federally funded research studies under procedures for handling scientific misconduct⁹ under federal regulations. These procedures contemplate researcher debarment for practices seriously deviating from those that are commonly accepted within the scientific community for proposing, conducting, or reporting research. Fines and compensation to individuals whose data have been improperly accessed or misused could provide additional deterrents to abuse.¹⁰ The research engine as a whole, however, should not be hindered by the remote possibility of an errant investigator reidentifying individuals without authorization. Several discussants suggested that discrimination or other harm caused by unauthorized reidentification of research participants seems largely to be a theoretical problem at this time. Moreover, in order to identify an individual in a pooled data set, a separate reference sample or at least some independent data about the individual is still generally required.

The second theme that emerged at the think tank takes a different, although not truly contradictory, approach to concerns regarding the risk of reidentification. This view is focused on the

Table 1 Highlights of individual state-of-the-science presentations

Presenter	Summary
Laura Lyman Rodriguez, PhD Keynote Speaker Office of Policy, Communication and Education, NHGRI/NIH	<p>Presentation: Is It or Isn't It? Evolving Policy Considerations Regarding Genomic Data and Identifiability</p> <p>Noted that in 2006 NHGRI sponsored a meeting that highlighted various definitions for "identifiable" (ref. 19)</p> <p>Reviewed past and current data-sharing policies</p> <p>Commented on the impact and availability of recent high-throughput sequencing technologies</p> <p>Emphasized the need for a proper balance between societal benefit and individual harm</p>
Misha Angrist, PhD, MS, MFA Keynote Speaker Duke Institute for Genome Sciences and Policy	<p>Presentation: Can You See the Real Me? Human Patients and Human Research Participants</p> <p>Provided insights from experience as an IRB member and participant in the personal genome project</p> <p>Proposed that issues with identifiability and privacy are directly a result of the research community and those overseeing human participant protections losing sight of the intended audience</p> <p>Proposed that deidentification should not be used to relieve researchers from their obligation to interact with their research participants</p> <p>Stressed the importance of the return of results for community engagement, transparency, and promotion of participant trust</p>
Bradley Malin, PhD, MS Health Information Privacy Laboratory at Vanderbilt University	<p>Presentation: "Omics" and the Changing Face of Identifiability</p> <p>Risk for reidentification is not unique to "omic" data; this was demonstrated more than a decade ago using demographic data</p> <p>Need to recognize that even with a link (e.g., matching sample), reidentification, although possible, is not probable</p> <p>Suggested privacy policy discussions should acknowledge that deidentification, although somewhat protective, is not a panacea and should focus on the context of risk, opportunities for harm associated with reidentification, and risk-mitigation strategies</p>
P. Pearl O'Rourke, MD Partners HealthCare	<p>Presentation: IRBs: Forced to Deal With "Identifiability" of Everything. A Daunting Mandate</p> <p>Identifiability of various data sets or specimens lies on a spectrum, but IRBs are forced to make a binary choice—"yes" or "no"—when determining whether a proposed study represents human subjects research</p> <p>Geneticists do not necessarily agree on whether particular types of "omic" data are identifiable</p> <p>Explored the potential impact of consent for all secondary research using biospecimens or genomic data</p>
Leslie Biesecker, MD Genetic Disease Research Branch, NHGRI	<p>Presentation: Hypothesis-Testing and Hypothesis-Generating Modes of Research</p> <p>Hypothesis-testing paradigms have been useful during decades of research using low-throughput technology</p> <p>Proposed a paradigm shift toward hypothesis-generating research, allowing investigators to assemble cohorts, collect molecular data, and use data to identify new phenotypes and generate new hypotheses</p> <p>Such 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</p>
John Wilbanks Ewing Marion Kauffman Foundation	<p>Presentation: Portable Legal Consent</p> <p>Data sharing can be a beneficial way for patients to affirm their autonomy and overcome feelings of vulnerability and disempowerment</p> <p>Currently there are no avenues for willing individuals to donate their health information to research studies</p> <p>Proposed the use of portable legal consent—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 and provides participants a comprehensive explanation of the scope of their consent</p>
Deborah Collyar Patient Advocates in Research	<p>Presentation: Making Sense of Genomics While Protecting People</p> <p>From the patient perspective, emphasized the need for plain language and clear information</p> <p>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</p> <p>Many patients want the biomedical research community to overcome intellectual property issues and share data</p>
Kenneth Chahine, PhD, JD Ancestry DNA	<p>Presentation: The Future of Genomic and Health Data</p> <p>The world is changing from one in which health data are locked away in physicians' offices and research databases to one in which individuals share large amounts of information through social networking and mobile phone applications</p> <p>Many Ancestry participants share their information to foster interaction</p> <p>Individuals' ideas about privacy could change when they see a personal benefit from sharing</p>
Jane Yakowitz Bambauer, JD University of Arizona	<p>Presentation: Seeding the Data Commons: Legal Safe Harbors for Research Data</p> <p>Argued that the risk for reidentification is likely overstated and cautioned against implementing protections when the credible downstream risks are still unknown and when the value of information sharing is great</p> <p>Proposed a three-step model that (i) includes a basic anonymization process, (ii) holds data producers immune from privacy-related liability, and (iii) ensures that malicious actors are held criminally liable for intentional reidentification and misuse of data</p>

IRB, institutional review board; NHGRI, National Human Genome Research Institute; NIH, National Institutes of Health.

Table 2 Outcomes of breakout group discussions

Working group	Outcome
<p>Group 1</p> <p>What factors should be considered in the development of a federal policy for access to publicly funded “omic” research data?</p>	<p>Outcomes of discussion: Proposed a pilot of a truly open-source data set that would allow access to all types of data, including “omic” data. Participants willing to contribute their biospecimens and data would be informed that all data would be made public and that anyone would be able to access them. Participants would be informed about the risks and benefits associated with such data sharing.</p> <p>The pilot would involve multiple data sets 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 data set would be a good start.</p> <p>The pilot could assemble a large cohort of individuals with specific phenotypes and track it longitudinally, with broad consents to overcome issues of controlled access.</p> <p>Patient advocacy groups, public health departments, and epidemiologists could, among others, serve as partners in recruiting individuals for such a cohort.</p>
<p>Group 2</p> <p>What considerations enter into determining whether “omic” data is identifiable?</p>	<p>Outcomes of discussion: There was consensus that all data are identifiable, but acknowledgment that reidentification 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. Therefore, the group proposed a three-tiered “Russian doll model.”</p> <p>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.</p> <p>The general controlled access tier (tier 2) allows investigators to apply for access to data that fall under 45 CFR 46.101b exemption 4 or data for which participants have consented to broad sharing.</p> <p>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.</p>
<p>Group 3</p> <p>What are appropriate ethical constraints to allowing researchers broad access to “omic” data?</p>	<p>Outcomes of discussion: 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.</p> <p>The burden of contact and recontact is not fully on the research community. Participants can be asked to maintain their correct contact information if they want to receive results.</p> <p>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.</p> <p>The group also discussed the need to include new technologies into current research processes, e.g., online consent that allows participants to choose the level of information they want to share.</p>
<p>Group 4</p> <p>How can society minimize any risks and maximize any participant benefits of “omic” research?</p>	<p>Outcomes of discussion: The levels of risk vary across data types. 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, and so forth.</p> <p>Real instances of misuse have related to issues of data security and, so far, reidentification has occurred primarily within the research setting. Therefore, the risks to research participants remain remote.</p> <p>Efforts to address remote risks could minimize the benefits to research participants and the community. Accordingly, studies should minimally report aggregate data.</p> <p>Streamlined data access committees should be used to control the release of data and investigators should be required to accept accountability by registering for access to data.</p> <p>In addition, training data are needed to develop and improve bioinformatics.</p>

concept of balancing privacy risk with the potential benefits of broad data sharing. There is developing consensus that anonymity is more difficult to guarantee in the genomic era.¹¹ Several meeting participants suggested that all “omic” data are theoretically identifiable. Moreover, the problem of reidentification is not unique to biomedical research, and existing public databases make reidentification of individuals easier.¹² Given our waning ability to guarantee complete deidentification of data, some experts felt that medical researchers need to improve communication with participants regarding the remote potential for data security breaches and counterbalance the risks by conveying the potential societal and individual benefits of data sharing. Such benefits include the medical utility of participant contributions of data and the ability to return clinically relevant incidental findings or other research results under appropriate circumstances.

DISCUSSION

As “omic” research advances and data are shared more broadly, communication of potential reidentification risks for participants is a challenge. Greater transparency is needed when informing participants about the limitations of data privacy, and efforts should be made to more clearly convey the complexity of this issue. Although consent forms routinely disclose that researchers cannot guarantee privacy, it is not clear whether research participants understand the rationale behind this limitation. In a study that examined comprehension of consent documents, participants scored poorly on a question evaluating the understanding of confidentiality issues.¹³ By improving communication with participants, they may become more willing to embrace sharing of their submitted data. Some meeting participants suggested that this is an opportunity to advance

recruitment in medical research. A recent European study suggests that participants may be less concerned about the risks of data sharing and more interested in potential positive outcomes of such sharing.¹⁴

Much discussion at the think tank focused on the utility and empowerment of data sharing and its potential to stimulate collaborative engagement with research participants. Individuals facing illness or chronic disease may visit and join Internet groups such as CancerCompass and PatientsLikeMe, where they can share medical outcomes and learn about the symptoms or treatment response of others. The growth of such electronic communication spaces indicates the importance attached by many to the benefits of sharing their personal medical data.

As society becomes more accustomed to sharing and receiving information through social media and mobile communications, people who commonly use these resources might be more willing to release their genomic information to researchers who have engendered their trust. Others who are less comfortable with, or have no access to, such commercial technologies may be less inclined to disclose their personal information broadly. Nonetheless, the increasing popularity of social media and mobile technologies may underscore a cultural shift toward greater openness about personal information extending beyond early adapters.

In a randomized trial of consent options for data sharing in genome research by McGuire *et al.*,¹⁵ 86.2% of research participants chose to release their genetic and clinical information in either open-access (through the Internet) or scientific (restricted for medical research only) databases, even after a debriefing and opportunity to change their data release option. A follow-up study showed that when deciding whether protecting privacy or advancing research is more important, participants who somewhat or highly trusted researchers predominantly chose advancing research as having greater priority.¹⁶ Participants in these studies were recruited by physicians at hospitals where they or their family members were receiving treatment, which may have influenced willingness to share data. The studies nonetheless suggest that research participants will incur some privacy risk with their individual genomic data when they trust those conducting the research and perceive that their information will benefit others.

Discussions of individual privacy risk in medical research trigger additional ethical challenges relating to risks for family members and genetically similar populations. If a privacy breach results in the disclosure of information pertaining to an inherited mutation, for example, that information may harm not only the individual research participant but also other genetically related people. Under such circumstances, an individual's choice to participate in research may result in group discrimination or stigmatization. A research participant belonging to a relatively homogeneous genetic population who permits broad sharing of individual research data may enable the discovery of incidental findings that impact the entire population. The severity of consequences for public trust and future research recruitment may, in such cases, be quite high even though the

statistical risk of reidentification of the individual research participant is quite low.

NIH data-sharing policy^{1,2} embraces the importance of data sharing as a tool for research progress, but we are only just beginning to explore how research participants understand and balance the risks and benefits of broad “omic” data sharing for translational medical research. Although extending researchers' access to “omic” data imposes a degree of individual and group privacy risks, some research participants might prefer the informational benefits of data sharing, particularly when offered the return of clinically actionable research results similar to the proposed collaborative informed cohort model.¹⁷ Others might prefer more conventional deidentification options. We would anticipate that when data-sharing strategies are clearly communicated during the consent process, and research participants understand the potential risks and benefits of available data-sharing options, they are likely to be more comfortable releasing their data for broad future research uses.

The NCI think tank concerning the identifiability of biospecimens and “omic” data not only motivated dialogue across a range of expert perspectives but also identified opportunities for agency guidance, policy development, and empirical research. As an initial next step, more thought must be given to use of the term “identifiability” in research documents, as well as to its application in regulations and guidance. Identifiability is often conflated with the term “privacy” in consent forms, although the meaning of these terms is not identical. Privacy is a multifaceted concept with roots in constitutional law, encompassing broadly the right to be free from unwarranted invasion of personal liberties and the ability to make personal decisions affecting marriage, contraception, and other intimate matters. Identifiability, on the other hand, refers simply to the ability to determine unique facts about somebody or to figure out who he or she is. Consent forms often speak to the ability to protect privacy, but in reality, investigators can only seek to protect identifiability.

Moreover, the term “identifiability” itself has multiple meanings. In the research context, the term refers to the technical capability to resolve individuals through linkage to matched samples or independent data. In the regulatory context, however, identifiability refers quite differently to whether specific investigators may readily ascertain a research participant's identity or gain access to specified identifiers. Therefore, “deidentified” data in the regulatory context may be scientifically identifiable, causing confusion when the term is used. Institutions that engage in “omic” research, including biorepositories and academic medical centers, should examine how their data protection policies reflect the different meanings of “identifiability” and ensure clarity in their policies and consent documents. Institutional guidelines should be developed to assist investigators and institutional review boards in determining whether specific “omic” data are or are not identifiable. Such guidelines should illustrate the connections among the various definitions, laws, and policies related to

privacy and identifiability. Beyond institutional policy, clarification regarding identifiability at a broader level can be implemented through the research industry, NIH policy measures, endorsements from professional organizations, and changes in regulations such as the Health Insurance Portability and Accountability Act and the DHHS Common Rule.

Additional important steps include the development of strategies for greater transparency in communicating data-sharing options and empirical investigation of participant preferences in balancing the risks and benefits of various data-sharing models. The public seems to support broad research uses of biospecimens and collected medical data,¹⁸ but the extent to which people understand the potential risks and benefits of different tiers of data sharing (i.e., controlled access versus public access) is not clear. To improve scientific literacy about biospecimen donation and possible modes of data sharing, a model suggested for consideration during the think tank discussion is that of organ donation. Organ donation programs provide education and community outreach, with the goal of enabling individuals to make informed decisions, documented on their driver's licenses, about donating organs for transplant. Similarly, we could adopt a community-based model to disclose the potential risks and benefits of donating biospecimens and associated data for medical research and allow individuals to document their choices on driver's licenses. Perhaps the most widely shared observation among think tank participants was the determination that additional empirical research about the perceived and actual risks and benefits of "omic" data sharing is needed. In particular, we should explore the underpinnings of information altruism (i.e., desire of participants to assist medical research by sharing their personal data broadly) and its relative impact across disease and population communities. Developing research collaborations with private corporations that provide personal genomic information can aid in moving the science forward at a more rapid pace, given that many of these companies have research arms and have expressed an interest in collaborating with academic institutions and the NIH. Such collaborations would need to proceed cautiously, however, in order to prevent untoward research focus on populations that have both the resources and the inclination to pay for personal genetic information. In addition to research into the relative risks versus benefits of data sharing, further statistical analysis of the probability that someone could be reidentified via his or her "omic" data is needed. This analysis should take into account fiduciary responsibilities and institutional relationships at various stages, including the consent process, storage and stewardship of data, and methods for addressing potential downstream misuse (although at present, misuse has not been documented as a widespread problem).

Another think tank suggestion was to conduct a pilot study of participants with a completely open-access data model and multiple types of "omic" or phenotypic data. Such a project could investigate the willingness of individuals to share data under this model, examine real-world consequences of truly

open-access data, assess the feasibility of enrolling participants using such a model, and describe the experiences of individuals and empirical evidence about whether data would be shared more rapidly. Results from such suggested research could guide the development of best practices for research institutions and biorepositories and inform research consent disclosures about informational risk.

In conclusion, the NCI think tank concerning the identifiability of biospecimens and "omic" data identified two main themes regarding how the research community should manage the difficulty of ensuring deidentification and the remote risk that future reidentification will occur and cause harm. The first theme concerned whether additional institutional or legal measures are necessary to address a violation of research participant privacy. The second concerned the extent to which greater understanding and expansion of data-sharing options might provide an opportunity to increase engagement of participants in the research enterprise. The NCI think tank also identified several opportunities for improved guidance, policy development, and empirical research. The challenge of deciphering different uses of the term "identifiability" was discussed, suggesting that more should be done to clarify the meaning for both researchers and study participants. The ability to communicate the changing reality of reidentification risks is fundamental to the development of best practices for seeking the consent of research participants in biospecimen collection and storage protocols. In addition, we must gather more empirical data about people's data-sharing preferences, to facilitate increased secondary research uses of their biospecimens and data and thereby hasten the pace of medical progress. Addressing these priorities should help the research community and policy makers in developing a measured approach to the issues of reidentification in the coming era of high-throughput population-based genomic research and should better define the balance between patient protections and advancing science.¹⁹

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at <http://www.nature.com/gim>

ACKNOWLEDGMENTS

Funds for the National Cancer Institute (NCI) think tank were provided by the Host Susceptibility Factors Branch (HSFB) and the Epidemiology and Genomics Research Program (EGRP) in the Division of Cancer Control and Population Sciences, NCI and the Biorepositories and Biospecimen Research Branch (BBRB), Division of Cancer Treatment and Diagnosis, NCI. The authors acknowledge the assistance with meeting planning of Elizabeth Gillanders (HSFB) and Jim Vaught (BBRB). The authors thank all of the participants of the NCI Identifiability think tank listed in **Supplementary Appendix S2** online for their intellectual contributions throughout the meeting.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. National Institutes of Health. NOT-OD-07-088: Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS). <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html> 2007.
2. National Institutes of Health. NOT-HG-10-006 Notice on Development of Data Sharing Policy for Sequence and Related Genomic Data. <http://grants.nih.gov/grants/guide/notice-files/NOT-HG-10-006.html> 2009.
3. Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 2008;4:e1000167.
4. Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. *Nat Genet* 2012;44:603–608.
5. Im Hae K, Gamazon ER, Nicolae DL, Cox NJ. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *Am J Hum Genet* 2012;90:591–598.
6. Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. *Science* 2004;305:183.
7. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science* 2013;339:321–324.
8. Department of Health and Human Services. Human subjects research protections: Enhancing protections for research subjects and reducing burden, delay and ambiguity for investigators. In: Department of Health and Human Services (ed). *Federal Register*, vol. 76; 2011: 44512–44531.
9. Department of Health and Human Services. Subpart A: responsibility of PHS Awardee and Applicant Institutions for dealing with and reporting possible misconduct in science. 45 CFR 50 Subpart A.
10. Presidential Commission for the Study of Bioethical Issues. Privacy and progress in whole genome sequencing. Department of Health and Human Services. Washington, DC, 2012. <http://bioethics.gov/cms/sites/default/files/PrivacyProgress5081.pdf>.
11. Schadt EE. The changing privacy landscape in the era of big data. *Mol Syst Biol* 2012;8:612.
12. Ohm P. Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA Law Rev* 2009;57:1701.
13. Ormond KE, Cirino AL, Helenowski IB, Chisholm RL, Wolf WA. Assessing the understanding of biobank participants. *Am J Med Genet A* 2009; 149A:188–198.
14. Hobbs A, Starkbaum J, Gottweis U, Wichmann HE, Gottweis H. The privacy-reciprocity connection in biobanking: comparing German with UK strategies. *Public Health Genomics* 2012;15:272–284.
15. McGuire AL, Oliver JM, Slashinski MJ, et al. To share or not to share: a randomized trial of consent for data sharing in genome research. *Genet Med* 2011;13:948–955.
16. Oliver JM, Slashinski MJ, Wang T, Kelly PA, Hilsenbeck SG, McGuire AL. Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. *Public Health Genomics* 2012;15:106–114.
17. Kohane IS, Mandl KD, Taylor PL, Holm IA, Nigrin DJ, Kunkel LM. Medicine. Reestablishing the researcher-patient compact. *Science* 2007;316: 836–837.
18. Trinidad SB, Fullerton SM, Ludman EJ, Jarvik GP, Larson EB, Burke W. Research ethics. Research practice and participant preferences: the growing gulf. *Science* 2011;331:287–288.
19. Lowrance WW, Collins FS. Ethics. Identifiability in genomic research. *Science* 2007;317:600–602.