Comments from Data Privacy Researchers

October 26, 2011

To:
The Department of Health and Human Services,
Office of the Secretary, and Food and Drug Administration

Re: Advance notice of proposed rulemaking: Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators, Docket ID number HHS–OPHS–2011–0005

Via Regulations.gov

To the Office of the Secretary, US Department of Health and Human Services:

As data privacy researchers, we appreciate the opportunity to comment on the Advance Notice of Proposed Rulemaking (ANPRM) on human subjects research protections that appeared in 76 Federal Register 44512 (July 26, 2011). These comments mostly address the fitness of HIPAA provisions in privacy issues raised by the ANPRM. Data privacy is emerging as a field, and data privacy is a cross-disciplinary pursuit, including researchers in policy, computer science, medical informatics, statistics, political science, and law. Our responses also support the submission made by Salil Vadhan at Harvard University, et al., which draws on recent advances in understanding data privacy from a theoretical computer science perspective.

Respectfully submitted,

Latanya Sweeney PhD
Director and Founder of the Data Privacy Lab
Harvard University
latanya@seas.harvard.edu

JOINED BY (ALPHABETICALLY)

Hal Abelson PhD
Professor Computer Science
Massachusetts Institute of Technology

Jonathan Aldrich PhD
Associate Professor
Carnegie Mellon University

Alessandro Acquisti PhD
Associate Professor
Carnegie Mellon University

Isa Bar-On PhD
Professor
Worcester Polytechnic Institute

Natasha Balac PhD
Super Computing Center
University of California San Diego

Elisa Bertino PhD
Professor Computer Science
Purdue University

Julia Brody PhD
Executive Director
Silent Spring Institute

Phil Brown PhD
Professor
Brown University

Chris Clifton PhD
Professor Computer Science
Purdue University

Wendy Chapman PhD
Associate Professor Biomedical Info.
University of California San Diego

Kamalika Chaudhuri PhD
Assistant Professor of Computer Science
University of California San Diego

Rui Chen
PhD Candidate Computer Science
Concordia University

Samuel Cheng PhD
Assistant Professor
University of Oklahoma

Electronic Privacy Information Center (EPIC)
Marc Rotenberg JD
www.epic.org/privacy/common-rule/

Lisa Fleischer PhD
Associate Professor of Computer Science
Dartmouth University

Ralph Gross PhD
Carnegie Mellon University

Tim A. Holt PhD MRCP FRCGP
Department of Primary Care
University of Oxford

Zhanglong Ji
Super Computer Center
University of California San Diego

Xiaoqian Jiang PhD
Center for Biomedical Informatics
University of California San Diego

Murat Kantarcioglu PhD
Associate Professor of Computer Science
University of Texas Dallas

Paul Kantor PhD
Distinguished Professor
Rutgers University

Katherine Kim, MPH, MBA
Professor in Residence of Biology
San Francisco State University

David Kotz
Professor Computer Science
Dartmouth University

Harry Lewis PhD
Professor Computer Science
Harvard University

Ben Livshits PhD
Microsoft Research

Danielle Mowery MS
Biomedical Informatics
University of Pittsburgh

Arvind Narayanan PhD
Stanford University

Joe Pato
Visiting Scientist
Massachusetts Institute of Technology

Patient Privacy Rights
Deborah Peel MD
http://patientprivacyrights.org/

Anand Sarwate PhD
Research Assistant Professor
Toyota Technological Institute at Chicago

Michael Shamos PhD
Distinguished Career Professor
Carnegie Mellon University

Vitaly Shmatikov PhD
Associate Professor of Computer Science
University of Texas Austin

Diane Strong
Professor
Worcester Polytechnic Institute

Kevin Sullivan PhD
Associate Professor of Computer Science
University of Virginia

Haixu Tang PhD
Associate Professor
Indiana University

Anthony Tomasic PhD
Director, VLIS
Carnegie Mellon University

Bengisu Tulu
Assistant Professor
Worcester Polytechnic Institute

Jaideep Vaidya PhD
Associate Professor
Rutgers Business School

Staal Vinterbo PhD
Associate Professor, Biomedical Informatics
University of California at San Diego

James Waldo PhD
Chief Technology Officer, Professor
Harvard University

Shuang Wang
PhD Student
University of Oklahoma

Gio Wiederhold PhD
Professor Emeritus
Stanford University

William Yasnoff MD PhD FACMI
NHII Advisors

Justin Zhan PhD
Associate Research Professor
Dakota State University

Executive Summary

Applying the HIPAA Privacy Rule standards for de-identification to research broadly in an attempt to protect against the informational risks described in the ANPRM *is poorly understood* and *all evidence suggests the HIPAA standards are gravely inadequate*. As examples, consider its lack of accountability and transparency in data sharing, the seeming lack of enforcement in light of the large number of allegations, HHS' own lack of demonstrated use, the proposed changes to the HIPAA Privacy Rule itself, the lack of a standard for its statistician provision, its lack of fitness to other kinds of data, including other forms of medical data beyond field-structured data, and the adverse impact that could result on sharing commercial data with researchers. Further, prohibiting re-identification, as posed by Question 63, would drive re-identification methods further into hidden, commercial activities and deprive the public, the research community and policy makers of knowledge about re-identification risks and potential harms to the public. Instead, what is needed is to invest in data privacy research and to establish channels for NCHS, NIST or a professional data privacy body to operationalize scientific research results so that real-world data-sharing decisions rely on the latest guidelines and best practices. Details for each of these points appears below and then, relevant parts are reiterated in specific response to questions 54, 55, 1, 63 and 64, in turn.

Lack of accountability and transparency in data sharing

The HIPAA Privacy Rule[1] was promulgated 9 years ago to protect patient privacy in the United States. Figure 1 shows data sharing before HIPAA and Figure 2 shows data sharing since HIPAA. Following the figures is a discussion of the sources used.

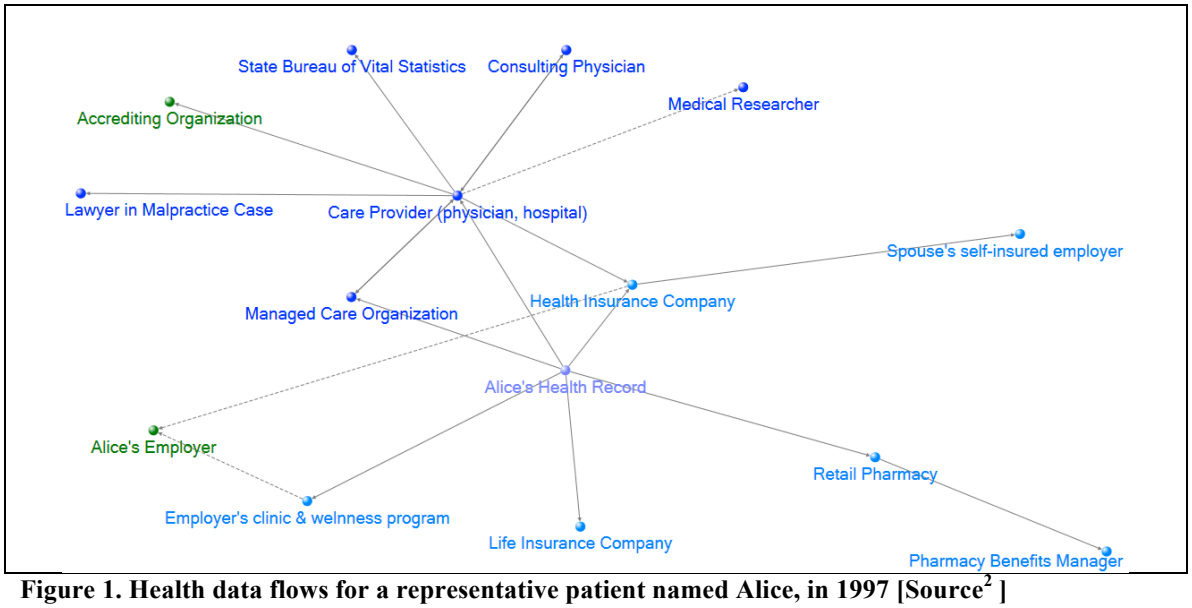

Figure 1. Health data flows for a representative patient named Alice, in 1997 [Source[2]]

[1] The Health Insurance Portability and Accountability Act (HIPAA) of 1996 (P.L.104-191)

[2] Clayton, P. et al. For the Record: Protecting Health Information. National Academy Press. 1997.http://www.nap.edu/catalog/5595.html

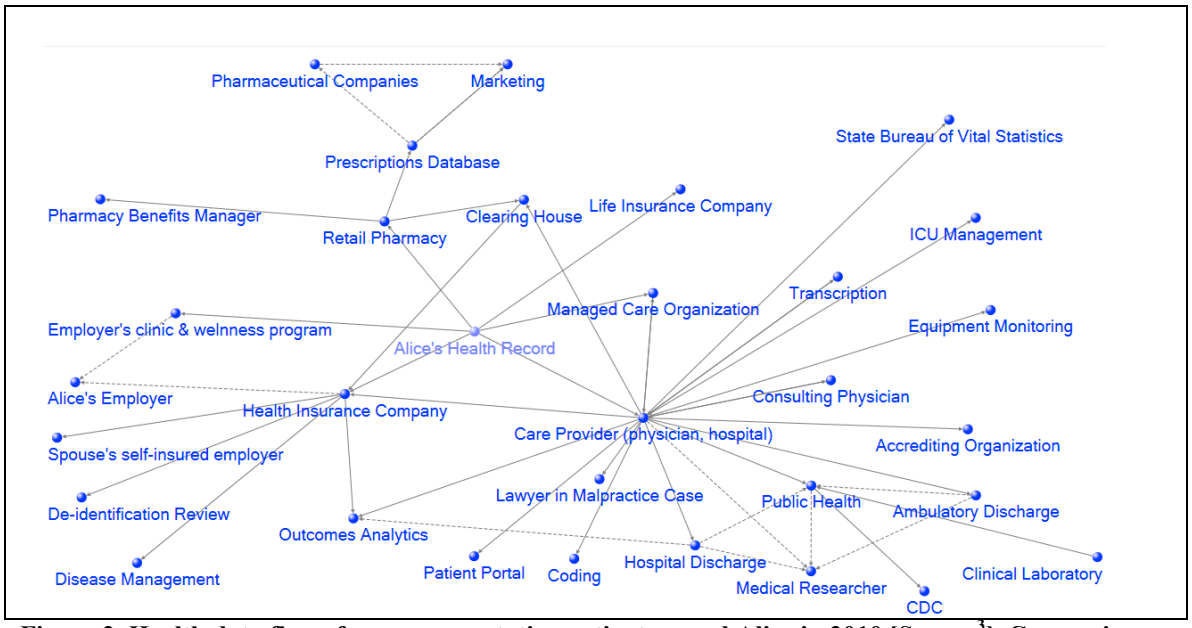

Figure 2. Health data flows for a representative patient named Alice in 2010 [Source[3]]. Comparing Figure 1 to Figure 2, the kinds of entities receiving information doubled, and today there is increased use of identifiable patient information and only long-term storage.

A committee from the National Research Council published a figure depicting flows of patient information about a hypothetical, but typical, patient named Alice.[4] Figure 1 is a reproduction, showing representative, not comprehensive, personal health data flows between organizations in 1997. The figure raised privacy concerns then because the sharing was hidden and because of a belief that greater data sharing increased risks of harms to patients.

Figure 2 shows representative flows of personal health data today. The number of entities receiving information more than doubled. New additions include data, outcome, and disease management organizations. There are more billing and offshore services. Entities receiving aggregate, temporary, or de-identified information now receive identifiable data stored long-term. Figure 2 shows results from a survey of the 6 year experience at the Data Privacy Lab at Carnegie Mellon University, researching patient data releases, de-identifying personal data, re-identifying ad hoc de-identifications, working on legal cases involving data identifiability, and advising government data efforts.[5] So, Figure 2 offers a description that is not even comprehensive.

The biggest problem is not more sharing, but patients and authorities having insufficient knowledge of sharing to assess harms and patients have no say. Expanding HIPAA standards to research broadly would similarly increase data sharing without researchers or research participants being able to assess harms.

---

[3] Data Privacy Lab, Carnegie Mellon University. September 30, 2010. http://dataprivacylab.org

[4] Clayton, P. et al. For the Record: Protecting Health Information. National Academy Press. 1997. http://www.nap.edu/catalog/5595.html

[5] Data Privacy Lab, Carnegie Mellon University. September 30, 2010. http://dataprivacylab.org

Lack of Enforcement and Large Number of Allegations

With so much data sharing, one expects to be able to point to a litany of harms, but a lack of enforcement and a lack of transparency confound findings. The Washington Post reported that the federal government received nearly 20,000 allegations of privacy violations under the Health Information and Portability and Accountability Act (HIPAA), but imposed no fines and prosecuted only two criminal cases by 2006.[6] As of 2010, there were 8 HIPAA criminal convictions[7] and a $1 million settlement with Rite-Aid[8]. Yet, in a 1996 survey of Fortune 500 companies, a third of the 84 respondents said they used medical records about employees to make hiring, firing and promotional decisions[9]). Allusions have been made to a banker crossing medical information with debtor information at his bank, and if a match results, tweaking creditworthiness accordingly[10]. True or not, it is certainly possible, and the lack of transparency in data sharing makes detection virtually impossible even though the harm can be egregious.

HHS' Own Lack of Demonstrated Use

Data considered sufficiently de-identified by the HIPAA Safe Harbor *can be freely used for any purpose* whatsoever, even published on the Internet. Yet, we are unaware of any publicly available data sets from the Centers for Medicare and Medicaid, the Centers for Disease Control and Prevention, or any other publicly available dataset available through the U.S. Department of Health and Human Services (HHS) that actually relies on the HIPAA Safe Harbor Provision. All publicly available datasets we found imposed additional redactions and sampling requirements.

For example, consider the Basic Stand Alone (BSA) Inpatient Public Use Files (PUF) named "CMS 2008 BSA Inpatient Claims PUF" with information from 2008 Medicare inpatient claims. This is a person-specific field-structured data file in which each record is an inpatient claim[11]. Beneficiaries have been selected as a 5% simple random sample (without replacement) from the approximately 48 million people eligible for Medicare at any time during 2008. Ages are given in 5-year age ranges and no residential geography is given; the patient resides somewhere in the United States. Additionally, a record for a sampled beneficiary is only included in a PUF if the combination of all analytic variables

---

[6] R Stein. Medical Privacy Law Nets No Fines: Lax Enforcement Puts Patients' Files At Risk, Critics Say. Washington Post. June 5, 2006. http://www.washingtonpost.com/wp-dyn/content/article/2006/06/04/AR2006060400672_pf.html

[7] Insider Threat Examples and 7th HIPAA Criminal Conviction. http://www.realtime-itcompliance.com/laws_regulations/2008/08/insider_threat_examples_7th_hi.htm

[8] Rite Aid Agrees to Pay $1 Million to Settle HIPAA Privacy Case as OCR Moves to Tighten Privacy Rules. Solutions Law Press. August 3, 2010 http://slphealthcareupdate.wordpress.com/2010/08/03/rite-aid-agrees-to-pay-1-million-to-settle-hipaa-privacy-case-as-ocr-moves-to-tighten-privacy-rules/

[9] D Linowes. "A Research Survey of Privacy in the Workplace," white paper available from the University of Illinois at Urbana-Champaign. (1996).

[10] B Woodward. The Computer-Based Patient Record and Confidentiality. N. Engl. J. Med. 333:1419-1422 (1995).

[11] CMS 2008 BSA Inpatient Claims PUF, http://www.cms.gov/BSAPUFS/03_Inpatient_Claims.asp

is shared by at least eleven (11) beneficiaries in the population (i.e., the dataset enforces k-anonymity[12], where *k*=11). In contrast, the HIPAA Safe Harbor Provision does not require sampling; the entire file could be released. Rather than 5-year age ranges, the year of birth is sufficient. Rather than the beneficiary being somewhere in the United States, the first 3- or 2-digit residential ZIP code can be given. And, there is no requirement to enforce k-anonymity. *Should the HIPAA Safe Harbor provision be tightened to actually reflect what HHS uses*?

The Office of the National Coordinator (ONC) in HHS recently conducted a re-identification experiment using data released under the Safe Harbor Provision and reported finding 2 re-identifications from 15,000 patients. The approach involved matching the de-identified data against identified commercial data on demographics and concluded that doing so "is much harder than expected"[13]. ONC and others seem to consider the test as evidence that the HIPAA Safe Harbor provision offers sufficient protection[14] even though the data that was the subject of the ONC re-identification test itself is not available publicly or even available for researchers to review or inspect or to test with other re-identification methodologies. In fact, HHS' own lack of sharing the test file that adhered to the HIPAA Safe Harbor provision undermines confidence in the standard and poses grave concerns about the validity and generalizability of HHS' findings.

If HHS itself does not rely on the HIPAA Safe Harbor provision when sharing data publicly, then it is difficult to consider encouraging others to use the HIPAA Safe Harbor provision broadly, for all forms of research data. Determining the adequacy of the HIPAA Safe Harbor provision is at best an evolving research effort, especially, given the rapidly changing landscape of our data-rich networked society. HHS should invest in data privacy research, support openness in sharing test data, encourage re-identification testing, and help establish channels for NCHS, NIST or a professional data privacy body to operationalize research results so that data sharing decisions and standards can rely on the latest guidelines and best practices.

Expansion of the HIPAA Privacy Rule Related to Breach Notices and Audit Logs

Anticipating increased data sharing due to widespread adoption of electronic health records, Congress strengthened HIPAA in the stimulus bill[15]. HHS has already proposed

---

[12] Sweeney L. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 557-570. http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.html.

[13] Kwok P and Lafky D. Harder than You Think: A Case Study of Re-identification Risk of HIPAA-Compliant Records. JSM 2011. http://dataprivacylab.org/projects/identifiability/kwokLafky.pdf

[14] El Emam K and Yakowitz J. Respondent Amici Brief. Sorrell v. IMS Health. U.S. Supreme Court. 2011. http://dataprivacylab.org/archives/sorrell/1.pdf

[15] The Health Information Technology for Economic and Clinical Health Act (HITECH Act) within the American Recovery and Reinvestment Act of 2009 (ARRA). Public Law 111 – 5. http://www.gpo.gov/fdsys/pkg/PLAW-111publ5/content-detail.html

requisite changes to HIPAA[16], leveraging breach laws and extending the use of audit logs. How effective are these? If HIPAA is adopted for all research use broadly, *how practical would breach laws and audit log requirements be*?

When information about thousands of patients is wrongfully released, breach laws require that the company notify the public of the number and nature of personal information disclosed. California officials received more than 800 reports of health data breaches in 5 months in 2009.[17] Privacy Rights Clearinghouse details 1,699 breaches involving more than 510 million personal records.[18] In a single breach, the U.S. Department of Veteran Affairs disclosed personal information on 26.5 million veterans, including their Social Security numbers, birth dates, and in some cases, health problems. Some breach laws require companies to notify people whose information was breached. Most breach laws protect companies from liability as an incentive for public announcement. Overall, breach notices tend to insulate companies from consequences of individual harms, and offer limited or no direct benefit to harmed individuals.

Audit logs record who accessed which patient's data and when the access occurred. The Los Angeles Times reports that an audit log records roughly 150 accesses from doctors, nurses, technicians, and billing clerks for at least part of a patient's health record during a hospital visit.[19] Hospitals have rotating staffs with dynamic role assignments, making it difficult to automatically identify inappropriate access at the time of occurrence, but in hindsight, audit logs can help. Audit logs documented hospital workers snooping at former President Clinton's record when he was undergoing heart surgery[20] and allegedly providing sensitive medical information about basketball player Kobe Bryant to a newspaper.[21] The first criminal conviction under HIPAA was an employee of a Seattle provider, who used the information to obtain credit cards in the patient's name.[22]

---

[16] Department of Health and Human Services. Proposed Modifications to the HIPAA Privacy, Security, and Enforcement Rules Under the Health Information Technology for Economic and Clinical Health Act. Federal Register. Vol 75 No. 134 July 14. 2010. http://edocket.access.gpo.gov/2010/pdf/2010-16718.pdf

[17] Dimick, C. Reports Pour in under California's New Privacy Laws. Journal of the American Health Information Management Association. Privacy and Security. July 7, 2009. http://journal.ahima.org/2009/07/07/cas-new-privacy-laws/

[18] Privacy Rights Clearinghouse. http://www.privacyrights.org/data-breach

[19] Health & Medicine (2006-06-26). "At risk of exposure: In the push for electronic health records, concern is growing about how well privacy can be safeguarded.". Los Angeles Times. http://articles.latimes.com/2006/jun/26/health/he-privacy26

[20] Stein, T. How Safe Are Your Computers. Hack Attack. Physicians Practice. February 1, 2005. http://www.physicianspractice.com/display/article/1462168/1588200

[21] Miller, M. Issues of Privacy in the Bryant Case. Los Angeles Times. September 8, 2003. http://articles.latimes.com/2003/sep/08/health/he-court8

[22] Tovino, S. U.S. Attorney applies HIPAA Criminal Penalty Provisions in First Conviction For Privacy Violations. August 27, 2004. http://www.law.uh.edu/healthlaw/perspectives/(ST)FirstHIPAAPrivacyConviction.pdf

Requiring breach reporting and audit logs would increase the expense of research and litigation risks, but the actual reduction in informational risks are not understood and technically-empowered alternatives could help, but have not been considered. [23], [24]

Lack of a Standard for the HIPAA Statistician Provision

The HIPAA Statistician Provision offers the ability to use risk assessment methodologies to determine whether any given data release has a "minimal risk of re-identification". Several strong approaches have come forward and others are being researched, but on its face, there are many shortcomings to this provision as currently written. How small is a "very small risk"? What qualifications should a person have to certify the results? What exactly are the criteria used to make the determination? HIPAA itself provides no answers, and so, *any two lay "statisticians" are allowed to make the determination, and in doing so, can give wildly different assessments and there are no external guidelines, and no required accountability or publication of the assessment criteria or finding*. What is needed is to invest in data privacy research and to establish channels for NCHS, NIST or a professional data privacy body to operationalize scientific results so that data-sharing decisions rely on the latest guidelines and best practices.

Under the HIPAA Statistician Provision, the risk for re-identification has to be "very small" but the regulation never provides any explicit means to quantify how small is very small. So, in fact, lawyers and statisticians alike were leery to use the provision. Sweeney introduced the Privacert Risk Assessment model for HIPAA Compliance ("Privacert Model") as a way of determining whether data are sufficiently de-identified under the HIPAA Statistician Provision.[25] The idea is simple: accept a dataset that does not make any more people identifiable than is made identifiable by the HIPAA Safe Harbor. As reported in earlier writings,[26] in general the identifiability of the HIPAA Safe Harbor is 0.04%, the exact value differs from state to state due to changes in population distributions and other publicly available datasets. The Privacert Model therefore, in general, accepts a dataset that may include fields not allowed by the HIPAA Safe Harbor (e.g., full dates and ZIP codes) provided no more people are put at risk to re-identification than would be allowed by the HIPAA Safe Harbor. The company Qunitles became the first to use a version of the Privacert approach in real-world practice after careful legal

---

[23] Sweeney L. "Weaving Innovative Privacy Technology into Fair Data Sharing Practices. Harvard Colloquium. Cambridge, MA October 2008. Video and/or slides available upon request.

[24] Sweeney L. Only You, Your Doctor, and Hundreds of Others Know. *under review* Manuscript available upon request.

[25] Sweeney, L. Data Sharing Under HIPAA: 12 Years Later. Invited presentation to the HHS Workshop on the HIPAA Privacy Rule's De-Identification Standard, Office of Civil Rights, U.S. Dept. of Health and Human Services, Washington, DC. March 8, 2010.
http://hhshipaaprivacy.com/assets/5/resources/Panel2_Sweeney.pdf

[26] Sweeney, L. Uniqueness of Simple Demographics in the U.S. Population. Carnegie Mellon University, School of Computer Science, Data Privacy Laboratory, Technical Report LIDAP-WP4. Pittsburgh: 2000. Shorter version available as: Simple Demographics Often Identify People Uniquely. Working Paper 2. 2000. http://dataprivacylab.org/projects/identifiability/index.html

and scientific review[27] and bioterrorism surveillance efforts sought to use the approach more widely. Over the last 7 years, numerous large insurance and data mining companies and government agencies have used the approach commercially.[28] Despite its use, however, there is no requirement that the Privacert model or any comparable techno-legal model be used.

In sharp contrast, the recent Supreme Court case, Sorrell v. IMS Health gave a glimpse at the lack of transparency and accountability currently afforded to data de-identified under the HIPAA Statistician provision when stronger models such as Privacert are not required.[29] IMS receives prescription data from pharmacies and sells versions of it to pharmaceutical companies for marketing purposes. The company relies on the HIPAA Statistician provision to receive data from pharmacies. Compliance is self-assessed. There is no external review of the company's de-identification process, no public detailed statement describing it, notwithstanding the years of litigation, and what is reported about it, exposes known vulnerabilities for re-identifying patients. Despite the growing explosion in data and data sharing over the past 8 years, the company seemingly did not seek less privacy-invasive approaches or to augment its approach with traditional remedies (e.g. Fair Information Practices or informed consent), and showed no interest in exploring new promising scientific or societal approaches to privacy protection. *Once data are deemed de-identified under HIPAA, under either the Safe Harbor provision or the Statistician provision, the data can be shared widely for any purpose*.

What is needed is to continue to invest in data privacy research and to establish channels for NCHS, NIST or a professional data privacy body to operationalize research results so that data sharing decisions rely on the latest guidelines and best practices. Doing so will not only improve data sharing practices but will also introduce many other forms of provable privacy protections.

Lack of Fitness for Other forms of Medical Data

De-identification provisions for the HIPAA Privacy Rule were designed narrowly with field-structured, person-specific claims data in mind. This perspective severely limits the ability of the provisions to apply to other forms of research data, even other forms of medical data. For example, clinical notes and letters between physicians are textual documents containing rich references to the lives of patients even when the Safe Harbor provisions are removed. For example, "this first occurred when she danced the lead to Showboat causing her to miss the first month" is the kind of references to employment

---

[27] Beach, J.Health Care Databases under HIPAA: Statistical Approaches to De-identification of Protected Health Information. DIMACS presentation. December 10, 2003. http://dimacs.rutgers.edu/Workshops/Health/abstracts.html and http://www.zurich.ibm.com/pdf/privacy/report3-final.pdf

[28] Privacert Risk Assessment Server (licensed to Privacert, Inc. by L. Sweeney, Carnegie Mellon University). http://privacert.com/assess/index.html

[29] Sweeney L. Patient Privacy Risks in U.S. Supreme Court Case Sorrell v. IMS Health Inc.: Response to Amici Brief of El Emam and Yakowitz. Data Privacy Lab Working Paper 1027-1015B. Cambridge 2011. http://dataprivacylab.org/projects/identifiability/1027.html

and lifestyle commonly occurring in clinical notes and physician letters.[30],[31] While often uniquely identifying, it does not require further redaction to be released in accordance with the HIPAA Safe Harbor provision. As another example, the HIPAA Safe Harbor provision requires dates to only reveal the year, and it does not impose any restrictions on transmission time or timestamps. So, consider a clinic that each day transmits a full day of events with time stamps from the previous day; even though the date reports only the year, one can infer the actual month, day and year of the events. Images and genomic information have problems too.

On the other hand, there have been scientific advances in ways to provide aggregate statistics, synthetic data, contingency tables, and other generalized knowledge with guarantees of anonymity e.g., [32] and [33], yet there is no incentive in HIPAA to use these approaches when practical because the HIPAA Safe Harbor provision allows more detailed data to be shared. As scientists develop more innovative remedies, there should be incentives established and distribution channels for NCHS, NIST or a professional data privacy body to operationalize research results so that real-world data sharing decisions rely on the latest guidelines and best practices.

Impact of Commercial Data Sharing on Researchers

HIPAA provisions were crafted from the perspective of governing the data source (e.g., hospital, physician, insurance company) and not the data recipient. Most researchers have historically been data sources, compiling information from observations, surveys and experiments, but increasingly, many researchers are no longer data collectors, but analyzers of data already collected. This fundamental shift places limits on the exposure to HIPAA litigation, criminal, and civil risks that researchers and research organizations may be willing to bear without seeking an alternative research structure that would not have such risk. *Research organizations that primarily rely on corporate data holders may form as a way of opting out of government imposed privacy oversight if HIPAA provisions are heavily imposed.*

Our understanding of ourselves is beginning to be transformed by computational social sciences and by genomics.[34] The reason is personal data: as we move through our lives we leave continuous, multifaceted digital traces that can be compiled into comprehensive

---

[30] Sweeney L. Replacing Personally-Identifying Information in Medical Records, the Scrub System. In: Cimino, JJ, ed. Proceedings, Journal of the American Medical Informatics Association (AMIA). Washington, DC: Hanley & Belfus, Inc, 1996:333-337. http://dataprivacylab.org/projects/scrub/index.html

[31] Clifton C et al. Anonymizing Textual Data and its Impact on Utility. http://projects.cerias.purdue.edu/TextAnon/

[32] Cynthia Dwork. Differential privacy: A survey of results. In Theory and Applications of Models of Computation, TAMC 2008, volume 4978, pages 1–19. Springer, 2008.

[33] Sweeney L. Demonstration of a Privacy-Preserving System that Performs an Unduplicated Accounting of Services across Homeless Programs. Data Privacy Lab Working Paper 902. Pittsburgh 2007, October 2008. http://dataprivacylab.org/projects/homeless/index2.html

[34] Lazer D, Pentland A, Adamic L et al. Computational Social Science. Science. 323(6). Feb 2009. pp.721-722.

pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.[35] Researchers who work in these emerging areas typically use data collected elsewhere, by corporations not bound to HIPAA or IRB regulation.

Many companies thrive through selling products and services that are enabled through the acquisition, curation and aggregation of personal data. For example, IMS Health collects personal prescription information from pharmacies and pharmacy benefits programs, and then uses it to sell market information to pharmaceutical companies.[36] Acxiom collects personal information from public records, such as marriage licenses and voter lists, and uses it to provide background checks.[37] Geisinger Health System, a large integrated health system, created a company called MedMining, which licenses its data to promote healthcare research, primarily to major pharmaceutical companies and large biotech companies.[38] Other companies (e.g. Google and Facebook) trade the use of online services for access to personal data.

Research access to commercial data can be unencumbered. For example, when Latanya Sweeney pioneered early research on finding and replacing personal information in textual clinical notes[39], she visited local area hospitals and left with data the same day, acquiring the data through the business office because her work was seen as a way to protect against possible litigation. In contrast, Peter Szolovits at MIT now reports spending 9 months of ongoing negotiations and delays to get the same kind of data from the same hospitals, impeding his research funded by the National Institutes of Health aimed at helping hospitals share data more freely with guarantees of patient anonymity.

Data Privacy, the field

Data Privacy is the study of risk and utility in data sharing arrangements. The question the discipline of Data Privacy seeks to answer is "For given data sharing arrangements, how can we construct integrated techno-policy systems that optimally minimize risk and maximize utility?" The discipline of Data Privacy may involve weaving traditional policy formations into existing technology, or creating innovative new technologies and policy altogether, thereby making it possible for Data Privacy to transcend the false belief that society must choose between privacy or utility, and instead pioneer new solutions so that society can enjoy both privacy and utility.

---

[35] Pentland A. Honest Signals: How They Shape Our World, Chapter 7, pp. 85-94, MIT Press, Cambridge, MA. 2008.

[36] IMS Health. IMS Facts at a Glance. As of September 30, 2010, http://www.imshealth.com/

[37] Acxiom. FAQs and EEOC Guidelines. As of September 30, 2010 http://www.acxiom.com/products_and_services/background_screening/faq/Pages/FAQs.aspx

[38] MedMining. Welcome to MedMining. As of September 30, 2010 http://www.medmining.com/

[39] Sweeney L. Replacing Personally-Identifying Information in Medical Records, the Scrub System. In: Cimino, JJ, ed. Proceedings, Journal of the American Medical Informatics Association (AMIA). Washington, DC: Hanley & Belfus, Inc, 1996:333-337. http://dataprivacylab.org/projects/scrub/index.html

Over the past 50 years the study of Data Privacy has grown from the efforts of a handful of statisticians exploring ways to render data anonymous and a handful of policy makers that largely ignored mathematical considerations when designing policies for sharing personal data widely, to an emerging broad cross-disciplinary field that has produced fundamental computational theories of anonymity, has designed algorithms for risk assessment and management, has introduced new policy approaches appropriate for a data rich networked society, and has spun off an industry of privacy technologies.

The de-identification provisions of the HIPAA Privacy Rule do not take advantage of advances in data privacy or the nuances it provides in terms of dealing with *different kinds of data* and finely *matching sensitivity to risk*. There needs to a channel for NCHS, NIST or a professional data privacy body to operationalize research results so that data sharing decisions rely on the latest guidelines and best practices. Doing so will not only improve data sharing practices but will also introduce many other forms of privacy options.

Finally, but perhaps most importantly, the proposed changes in the ANPRM (Question 63) threatens to weaken data privacy as a field by prohibiting re-identification. This could further *drive re-identification into hidden, commercial activities and deprive the public, the research community and policy makers of knowledge about re-identification risks and potential harms to the public*. Understanding risks to re-identification are important to understanding scientific privacy remedies.

Prohibiting re-identification for data privacy research is really bad because:

(1) It limits enforcement and accountability because data can be released and vulnerabilities found but researchers would be prohibited from blowing the whistle.
(2) It ensures obsoleteness, as we will not be able to determine when current methods are insufficient.
(3) it impairs the development of better methods that provide utility and privacy.

For convenience, the following pages reiterate parts of this summary specific to the questions 1, 54, 55, 63 and 64, in turn.

Question 1. *Is the current definition of ''minimal risk'' in the regulations (45 CFR 46.102(i)—research activities where ''the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests'')—appropriate? If not, how should it be changed?*

Response:

Relevant parts of the Executive Summary appear below. Responses to questions 54, 55, 63, and 64 appear thereafter.

Because several other questions relate to HIPAA as a means of determining "minimal risk", our response to this question focuses on the determination of minimal risk in HIPAA.

Under the HIPAA Statistician Provision, the notion of "minimal risk" is operationalized as the risk for re-identification being "very small". The regulation does not state that the risk needs to be compared to risks of daily life. In fact, there are many shortcomings to this provision as currently written. How small is a "very small risk"? What qualifications should a person have to certify the results? What exactly are the criteria used to make the determination? HIPAA itself provides no answers, and so, *any two lay "statisticians" are allowed to make the determination, and in doing so, can give wildly different assessments and there are no external guidelines, and no required accountability or publication of the assessment criteria or finding*. What is needed is to invest in data privacy research and to establish channels for NCHS, NIST or a professional data privacy body to operationalize scientific results so that data-sharing decisions rely on the latest guidelines and best practices.

Because of its lack of specificity, lawyers and statisticians alike were leery to use the provision. Sweeney introduced the Privacert Risk Assessment model for HIPAA Compliance ("Privacert Model") as a way of determining whether data are sufficiently de-identified under the HIPAA Statistician Provision.[40] The idea is simple: accept a dataset that does not make any more people identifiable than is made identifiable by the HIPAA Safe Harbor. As reported in earlier writings,[41] in general the identifiability of the HIPAA Safe Harbor is 0.04%, the exact value differs from state to state due to changes in population distributions and other publicly available datasets. The Privacert Model therefore, in general, accepts a dataset that may include fields not allowed by the HIPAA Safe Harbor (e.g., full dates and ZIP codes) provided no more people are put at risk to re-

[40] Sweeney, L. Data Sharing Under HIPAA: 12 Years Later. Invited presentation to the HHS Workshop on the HIPAA Privacy Rule's De-Identification Standard, Office of Civil Rights, U.S. Dept. of Health and Human Services, Washington, DC. March 8, 2010.
http://hhshipaaprivacy.com/assets/5/resources/Panel2_Sweeney.pdf

[41] Sweeney, L. Uniqueness of Simple Demographics in the U.S. Population. Carnegie Mellon University, School of Computer Science, Data Privacy Laboratory, Technical Report LIDAP-WP4. Pittsburgh: 2000. Shorter version available as: Simple Demographics Often Identify People Uniquely. Working Paper 2. 2000. http://dataprivacylab.org/projects/identifiability/index.html

identification than would be allowed by the HIPAA Safe Harbor. The company Qunitles became the first to use a version of the Privacert approach in real-world practice after careful legal and scientific review[42] and bioterrorism surveillance efforts sought to use the approach more widely. Over the last 7 years, numerous large insurance and data mining companies and government agencies have used the approach commercially.[43] Despite its use, however, there is no requirement that the Privacert model or any comparable techno-legal model be used.

In sharp contrast, the recent Supreme Court case, Sorrell v. IMS Health gave a glimpse at the lack of transparency and accountability currently afforded to data de-identified under the HIPAA Statistician provision when stronger models such as Privacert are not required.[44] IMS receives prescription data from pharmacies and sells versions of it to pharmaceutical companies for marketing purposes. The company relies on the HIPAA Statistician provision to receive data from pharmacies. Compliance is self-assessed. There is no external review of the company's de-identification process, no public detailed statement describing it, notwithstanding the years of litigation, and what is reported about it, exposes known vulnerabilities for re-identifying patients. Despite the growing explosion in data and data sharing over the past 8 years, the company seemingly did not seek less privacy-invasive approaches or to augment its approach with traditional remedies (e.g. Fair Information Practices or informed consent), and showed no interest in exploring new promising scientific or societal approaches to privacy protection. *Once data are deemed de-identified under HIPAA, under either the Safe Harbor provision or the Statistician provision, the data can be shared widely for any purpose.*

Over the past 50 years the study of Data Privacy has grown from the efforts of a handful of statisticians exploring ways to render data anonymous and a handful of policy makers that largely ignored mathematical considerations when designing policies for sharing personal data widely, to an emerging broad cross-disciplinary field that has produced fundamental computational theories of anonymity, has designed algorithms for risk assessment and management, has introduced new policy approaches appropriate for a data rich networked society, and has spun off an industry of privacy technologies.

Data Privacy is the study of risk and utility in data sharing arrangements. The question the discipline of Data Privacy seeks to answer is "For given data sharing arrangements, how can we construct integrated techno-policy systems that optimally minimize risk and maximize utility?" The discipline of Data Privacy may involve weaving traditional policy formations into existing technology, or creating innovative new technologies and policy

---

[42] Beach, J.Health Care Databases under HIPAA: Statistical Approaches to De-identification of Protected Health Information. DIMACS presentation. December 10, 2003. http://dimacs.rutgers.edu/Workshops/Health/abstracts.html and http://www.zurich.ibm.com/pdf/privacy/report3-final.pdf

[43] Privacert Risk Assessment Server (licensed to Privacert, Inc. by L. Sweeney, Carnegie Mellon University). http://privacert.com/assess/index.html

[44] Sweeney L. Patient Privacy Risks in U.S. Supreme Court Case Sorrell v. IMS Health Inc.: Response to Amici Brief of El Emam and Yakowitz. Data Privacy Lab Working Paper 1027-1015B. Cambridge 2011. http://dataprivacylab.org/projects/identifiability/1027.html

altogether, thereby making it possible for Data Privacy to transcend the false belief that society must choose between privacy or utility, and instead pioneer new solutions so that society can enjoy both privacy and utility. By utility we mean the benefits, usefulness and profits made possible by the data sharing arrangement. By risk we mean the possibility the data sharing arrangement may result in an explicit privacy violation or a harm, including economic harms to the data subject.

What is needed is to continue to invest in data privacy research and to establish channels for NCHS, NIST or a professional data privacy body to operationalize research results so that data sharing decisions rely on the latest guidelines and best practices. Doing so will not only improve data sharing practices but will also introduce many other forms of provable privacy protections.

-- next question starts on next page --

Question 54. *Will use of the HIPAA Privacy Rule's standards for identifiable and de-identified information, and limited data sets, facilitate the implementation of the data security and information protection provisions being considered? Are the HIPAA standards, which were designed for dealing with health information, appropriate for use in all types of research studies, including social and behavioral research? If the HIPAA standards are not appropriate for all studies, what standards would be more appropriate?*

Response:

Most of the Executive Summary responds to this question, so it is reprinted below over the next 11 pages. Responses to questions 55, 63, and 64 appear thereafter.

Applying the HIPAA Privacy Rule standards for de-identification to research broadly in an attempt to protect against the informational risks described in the ANPRM *is poorly understood* and *all evidence suggests the HIPAA standards are gravely inadequate*. As examples, consider its lack of accountability and transparency in data sharing, the seeming lack of enforcement in light of the large number of allegations, HHS' own lack of demonstrated use, the proposed changes to the HIPAA Privacy Rule itself, the lack of a standard for its statistician provision, its lack of fitness to other kinds of data, including other forms of medical data beyond field-structured data, and the adverse impact that could result on sharing commercial data with researchers. Further, prohibiting re-identification, as posed by Question 63, would drive re-identification methods further into hidden, commercial activities and deprive the public, the research community and policy makers of knowledge about re-identification risks and potential harms to the public. Instead, what is needed is to invest in data privacy research and to establish channels for NCHS, NIST or a professional data privacy body to operationalize scientific research results so that real-world data-sharing decisions rely on the latest guidelines and best practices. Details for each of these points appears below and then, relevant parts are reiterated in specific response to questions 54, 55, 1, 63 and 64, in turn.

Lack of accountability and transparency in data sharing

The HIPAA Privacy Rule[45] was promulgated 9 years ago to protect patient privacy in the United States. Figure 1 shows data sharing before HIPAA and Figure 2 shows data sharing since HIPAA. Following the figures is a discussion of the sources used.

[45] The Health Insurance Portability and Accountability Act (HIPAA) of 1996 (P.L.104-191)

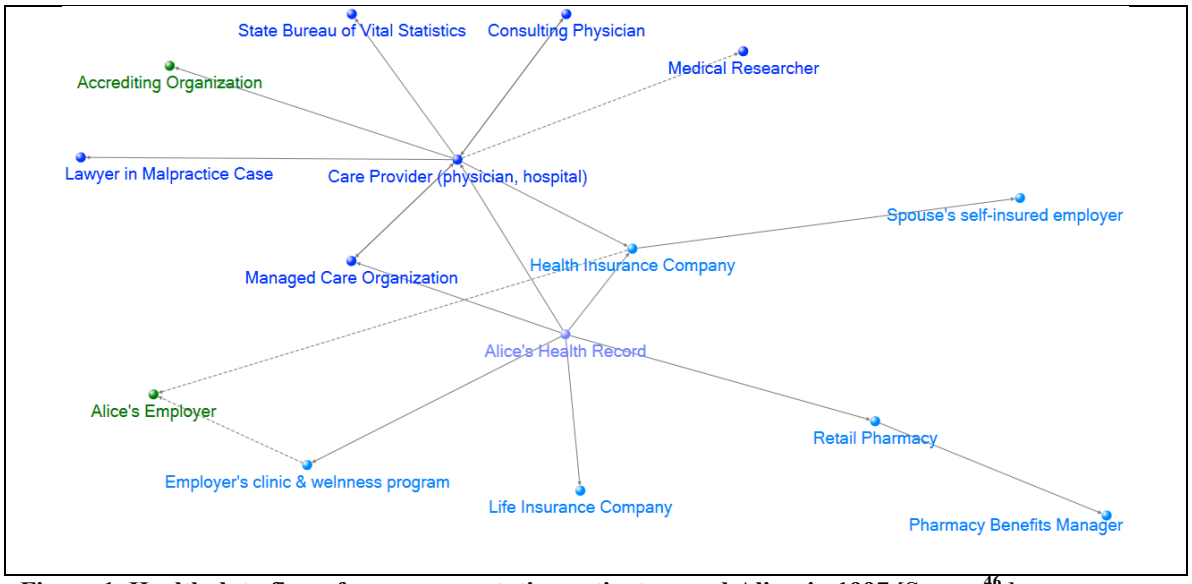

Figure 1. Health data flows for a representative patient named Alice, in 1997 [Source[46]]

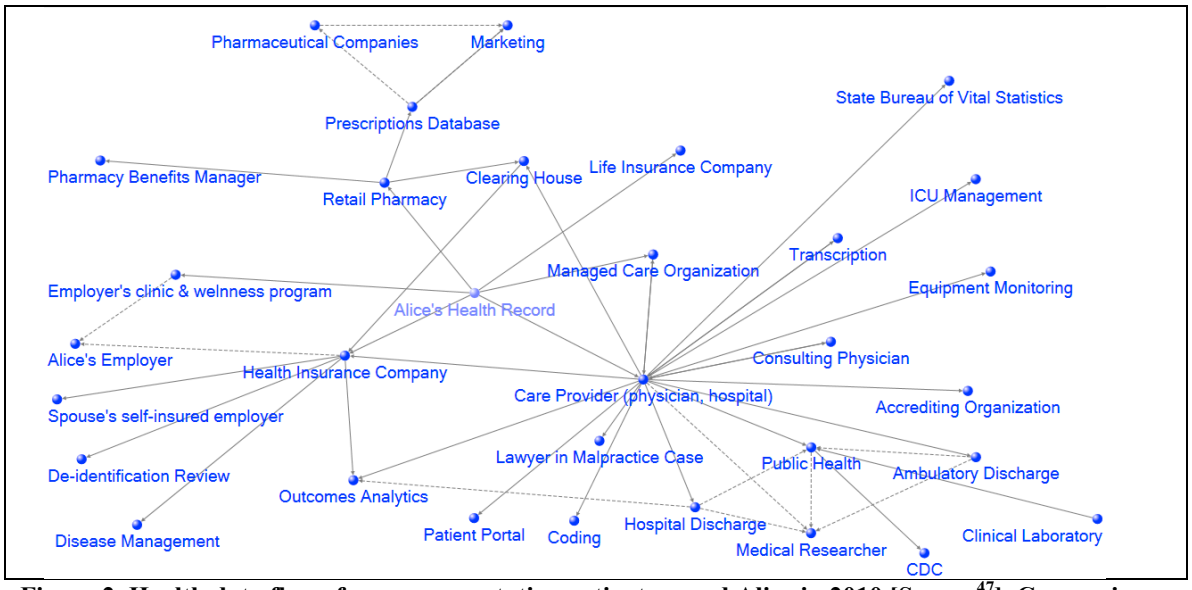

Figure 2. Health data flows for a representative patient named Alice in 2010 [Source[47]]. Comparing Figure 1 to Figure 2, the kinds of entities receiving information doubled, and today there is increased use of identifiable patient information and only long-term storage.

[46] Clayton, P. et al. For the Record: Protecting Health Information. National Academy Press. 1997.http://www.nap.edu/catalog/5595.html

[47] Data Privacy Lab, Carnegie Mellon University. September 30, 2010. http://dataprivacylab.org

A committee from the National Research Council published a figure depicting flows of patient information about a hypothetical, but typical, patient named Alice.[48] Figure 1 is a reproduction, showing representative, not comprehensive, personal health data flows between organizations in 1997. The figure raised privacy concerns then because the sharing was hidden and because of a belief that greater data sharing increased risks of harms to patients.

Figure 2 shows representative flows of personal health data today. The number of entities receiving information more than doubled. New additions include data, outcome, and disease management organizations. There are more billing and offshore services. Entities receiving aggregate, temporary, or de-identified information now receive identifiable data stored long-term. Figure 2 shows results from a survey of the 6 year experience at the Data Privacy Lab at Carnegie Mellon University, researching patient data releases, de-identifying personal data, re-identifying ad hoc de-identifications, working on legal cases involving data identifiability, and advising government data efforts.[49] So, Figure 2 offers a description that is not even comprehensive.

The biggest problem is not more sharing, but patients and authorities having insufficient knowledge of sharing to assess harms and patients have no say. Expanding HIPAA standards to research broadly would similarly increase data sharing without researchers or research participants being able to assess harms.

Lack of Enforcement and Large Number of Allegations

With so much data sharing, one expects to be able to point to a litany of harms, but a lack of enforcement and a lack of transparency confound findings. The Washington Post reported that the federal government received nearly 20,000 allegations of privacy violations under the Health Information and Portability and Accountability Act (HIPAA), but imposed no fines and prosecuted only two criminal cases by 2006.[50] As of 2010, there were 8 HIPAA criminal convictions[51] and a $1 million settlement with Rite-Aid[52]. Yet, in a 1996 survey of Fortune 500 companies, a third of the 84 respondents said they used medical records about employees to make hiring, firing and promotional decisions[53]). Allusions have been made to a banker crossing medical information with debtor

---

[48] Clayton, P. et al. For the Record: Protecting Health Information. National Academy Press. 1997. http://www.nap.edu/catalog/5595.html

[49] Data Privacy Lab, Carnegie Mellon University. September 30, 2010. http://dataprivacylab.org

[50] R Stein. Medical Privacy Law Nets No Fines: Lax Enforcement Puts Patients' Files At Risk, Critics Say. Washington Post. June 5, 2006. http://www.washingtonpost.com/wp-dyn/content/article/2006/06/04/AR2006060400672_pf.html

[51] Insider Threat Examples and 7th HIPAA Criminal Conviction. http://www.realtime-itcompliance.com/laws_regulations/2008/08/insider_threat_examples_7th_hi.htm

[52] Rite Aid Agrees to Pay $1 Million to Settle HIPAA Privacy Case as OCR Moves to Tighten Privacy Rules. Solutions Law Press. August 3, 2010 http://slphealthcareupdate.wordpress.com/2010/08/03/rite-aid-agrees-to-pay-1-million-to-settle-hipaa-privacy-case-as-ocr-moves-to-tighten-privacy-rules/

[53] D Linowes. "A Research Survey of Privacy in the Workplace," white paper available from the University of Illinois at Urbana-Champaign. (1996).

information at his bank, and if a match results, tweaking creditworthiness accordingly[54]. True or not, it is certainly possible, and the lack of transparency in data sharing makes detection virtually impossible even though the harm can be egregious.

HHS' Own Lack of Demonstrated Use

Data considered sufficiently de-identified by the HIPAA Safe Harbor *can be freely used for any purpose* whatsoever, even published on the Internet. Yet, we are unaware of any publicly available data sets from the Centers for Medicare and Medicaid, the Centers for Disease Control and Prevention, or any other publicly available dataset available through the U.S. Department of Health and Human Services (HHS) that actually relies on the HIPAA Safe Harbor Provision. All publicly available datasets we found imposed additional redactions and sampling requirements.

For example, consider the Basic Stand Alone (BSA) Inpatient Public Use Files (PUF) named "CMS 2008 BSA Inpatient Claims PUF" with information from 2008 Medicare inpatient claims. This is a person-specific field-structured data file in which each record is an inpatient claim[55]. Beneficiaries have been selected as a 5% simple random sample (without replacement) from the approximately 48 million people eligible for Medicare at any time during 2008. Ages are given in 5-year age ranges and no residential geography is given; the patient resides somewhere in the United States. Additionally, a record for a sampled beneficiary is only included in a PUF if the combination of all analytic variables is shared by at least eleven (11) beneficiaries in the population (i.e., the dataset enforces k-anonymity[56], where k=11). In contrast, the HIPAA Safe Harbor Provision does not require sampling; the entire file could be released. Rather than 5-year age ranges, the year of birth is sufficient. Rather than the beneficiary being somewhere in the United States, the first 3- or 2-digit residential ZIP code can be given. And, there is no requirement to enforce k-anonymity. *Should the HIPAA Safe Harbor provision be tightened to actually reflect what HHS uses*?

The Office of the National Coordinator (ONC) in HHS recently conducted a re-identification experiment using data released under the Safe Harbor Provision and reported finding 2 re-identifications from 15,000 patients. The approach involved matching the de-identified data against identified commercial data on demographics and concluded that doing so "is much harder than expected"[57]. ONC and others seem to consider the test as evidence that the HIPAA Safe Harbor provision offers sufficient

---

[54] B Woodward. The Computer-Based Patient Record and Confidentiality. N. Engl. J. Med. 333:1419-1422 (1995).

[55] CMS 2008 BSA Inpatient Claims PUF, http://www.cms.gov/BSAPUFS/03_Inpatient_Claims.asp

[56] Sweeney L. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 557-570. http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.html.

[57] Kwok P and Lafky D. Harder than You Think: A Case Study of Re-identification Risk of HIPAA-Compliant Records. JSM 2011. http://dataprivacylab.org/projects/identifiability/kwokLafky.pdf

protection[58] even though the data that was the subject of the ONC re-identification test itself is not available publicly or even available for researchers to review or inspect or to test with other re-identification methodologies. In fact, HHS' own lack of sharing the test file that adhered to the HIPAA Safe Harbor provision undermines confidence in the standard and poses grave concerns about the validity and generalizability of HHS' findings.

If HHS itself does not rely on the HIPAA Safe Harbor provision when sharing data publicly, then it is difficult to consider encouraging others to use the HIPAA Safe Harbor provision broadly, for all forms of research data. Determining the adequacy of the HIPAA Safe Harbor provision is at best an evolving research effort, especially, given the rapidly changing landscape of our data-rich networked society. HHS should invest in data privacy research, support openness in sharing test data, encourage re-identification testing, and help establish channels for NCHS, NIST or a professional data privacy body to operationalize research results so that data sharing decisions and standards can rely on the latest guidelines and best practices.

Expansion of the HIPAA Privacy Rule Related to Breach Notices and Audit Logs

Anticipating increased data sharing due to widespread adoption of electronic health records, Congress strengthened HIPAA in the stimulus bill[59]. HHS has already proposed requisite changes to HIPAA[60], leveraging breach laws and extending the use of audit logs. How effective are these? If HIPAA is adopted for all research use broadly, *how practical would breach laws and audit log requirements be*?

When information about thousands of patients is wrongfully released, breach laws require that the company notify the public of the number and nature of personal information disclosed. California officials received more than 800 reports of health data breaches in 5 months in 2009.[61] Privacy Rights Clearinghouse details 1,699 breaches involving more than 510 million personal records.[62] In a single breach, the U.S. Department of Veteran Affairs disclosed personal information on 26.5 million veterans, including their Social Security numbers, birth dates, and in some cases, health problems. Some breach laws require companies to notify people whose information was breached. Most breach laws protect companies from liability as an incentive for public announcement. Overall, breach

---

[58] El Emam K and Yakowitz J. Respondent Amici Brief. Sorrell v. IMS Health. U.S. Supreme Court. 2011. http://dataprivacylab.org/archives/sorrell/1.pdf

[59] The Health Information Technology for Economic and Clinical Health Act (HITECH Act) within the American Recovery and Reinvestment Act of 2009 (ARRA). Public Law 111 – 5. http://www.gpo.gov/fdsys/pkg/PLAW-111publ5/content-detail.html

[60] Department of Health and Human Services. Proposed Modifications to the HIPAA Privacy, Security, and Enforcement Rules Under the Health Information Technology for Economic and Clinical Health Act. Federal Register. Vol 75 No. 134 July 14. 2010. http://edocket.access.gpo.gov/2010/pdf/2010-16718.pdf

[61] Dimick, C. Reports Pour in under California's New Privacy Laws. Journal of the American Health Information Management Association. Privacy and Security. July 7, 2009. http://journal.ahima.org/2009/07/07/cas-new-privacy-laws/

[62] Privacy Rights Clearinghouse. http://www.privacyrights.org/data-breach

notices tend to insulate companies from consequences of individual harms, and offer limited or no direct benefit to harmed individuals.

Audit logs record who accessed which patient's data and when the access occurred. The Los Angeles Times reports that an audit log records roughly 150 accesses from doctors, nurses, technicians, and billing clerks for at least part of a patient's health record during a hospital visit.[63] Hospitals have rotating staffs with dynamic role assignments, making it difficult to automatically identify inappropriate access at the time of occurrence, but in hindsight, audit logs can help. Audit logs documented hospital workers snooping at former President Clinton's record when he was undergoing heart surgery[64] and allegedly providing sensitive medical information about basketball player Kobe Bryant to a newspaper.[65] The first criminal conviction under HIPAA was an employee of a Seattle provider, who used the information to obtain credit cards in the patient's name.[66]

Requiring breach reporting and audit logs would increase the expense of research and litigation risks, but the actual reduction in informational risks are not understood and technically-empowered alternatives could help, but have not been considered. [67], [68]

Lack of a Standard for the HIPAA Statistician Provision

The HIPAA Statistician Provision offers the ability to use risk assessment methodologies to determine whether any given data release has a "minimal risk of re-identification". Several strong approaches have come forward and others are being researched, but on its face, there are many shortcomings to this provision as currently written. How small is a "very small risk"? What qualifications should a person have to certify the results? What exactly are the criteria used to make the determination? HIPAA itself provides no answers, and so, *any two lay "statisticians" are allowed to make the determination, and in doing so, can give wildly different assessments and there are no external guidelines, and no required accountability or publication of the assessment criteria or finding*. What is needed is to invest in data privacy research and to establish channels for NCHS, NIST or a professional data privacy body to operationalize scientific results so that data-sharing decisions rely on the latest guidelines and best practices.

---

[63] Health & Medicine (2006-06-26). "At risk of exposure: In the push for electronic health records, concern is growing about how well privacy can be safeguarded.". Los Angeles Times. http://articles.latimes.com/2006/jun/26/health/he-privacy26

[64] Stein, T. How Safe Are Your Computers. Hack Attack. Physicians Practice. February 1, 2005. http://www.physicianspractice.com/display/article/1462168/1588200

[65] Miller, M. Issues of Privacy in the Bryant Case. Los Angeles Times. September 8, 2003. http://articles.latimes.com/2003/sep/08/health/he-court8

[66] Tovino, S. U.S. Attorney applies HIPAA Criminal Penalty Provisions in First Conviction For Privacy Violations. August 27, 2004. http://www.law.uh.edu/healthlaw/perspectives/(ST)FirstHIPAAPrivacyConviction.pdf

[67] Sweeney L. "Weaving Innovative Privacy Technology into Fair Data Sharing Practices. Harvard Colloquium. Cambridge, MA October 2008. Video and/or slides available upon request.

[68] Sweeney L. Only You, Your Doctor, and Hundreds of Others Know. *under review* Manuscript available upon request.

Under the HIPAA Statistician Provision, the risk for re-identification has to be "very small" but the regulation never provides any explicit means to quantify how small is very small. So, in fact, lawyers and statisticians alike were leery to use the provision. Sweeney introduced the Privacert Risk Assessment model for HIPAA Compliance ("Privacert Model") as a way of determining whether data are sufficiently de-identified under the HIPAA Statistician Provision.[69] The idea is simple: accept a dataset that does not make any more people identifiable than is made identifiable by the HIPAA Safe Harbor. As reported in earlier writings,[70] in general the identifiability of the HIPAA Safe Harbor is 0.04%, the exact value differs from state to state due to changes in population distributions and other publicly available datasets. The Privacert Model therefore, in general, accepts a dataset that may include fields not allowed by the HIPAA Safe Harbor (e.g., full dates and ZIP codes) provided no more people are put at risk to re-identification than would be allowed by the HIPAA Safe Harbor. The company Qunitles became the first to use a version of the Privacert approach in real-world practice after careful legal and scientific review[71] and bioterrorism surveillance efforts sought to use the approach more widely. Over the last 7 years, numerous large insurance and data mining companies and government agencies have used the approach commercially.[72] Despite its use, however, there is no requirement that the Privacert model or any comparable techno-legal model be used.

In sharp contrast, the recent Supreme Court case, Sorrell v. IMS Health gave a glimpse at the lack of transparency and accountability currently afforded to data de-identified under the HIPAA Statistician provision when stronger models such as Privacert are not required.[73] IMS receives prescription data from pharmacies and sells versions of it to pharmaceutical companies for marketing purposes. The company relies on the HIPAA Statistician provision to receive data from pharmacies. Compliance is self-assessed. There is no external review of the company's de-identification process, no public detailed statement describing it, notwithstanding the years of litigation, and what is reported about it, exposes known vulnerabilities for re-identifying patients. Despite the growing explosion in data and data sharing over the past 8 years, the company seemingly did not

---

[69] Sweeney, L. Data Sharing Under HIPAA: 12 Years Later. Invited presentation to the HHS Workshop on the HIPAA Privacy Rule's De-Identification Standard, Office of Civil Rights, U.S. Dept. of Health and Human Services, Washington, DC. March 8, 2010. http://hhshipaaprivacy.com/assets/5/resources/Panel2_Sweeney.pdf

[70] Sweeney, L. Uniqueness of Simple Demographics in the U.S. Population. Carnegie Mellon University, School of Computer Science, Data Privacy Laboratory, Technical Report LIDAP-WP4. Pittsburgh: 2000. Shorter version available as: Simple Demographics Often Identify People Uniquely. Working Paper 2. 2000. http://dataprivacylab.org/projects/identifiability/index.html

[71] Beach, J.Health Care Databases under HIPAA: Statistical Approaches to De-identification of Protected Health Information. DIMACS presentation. December 10, 2003. http://dimacs.rutgers.edu/Workshops/Health/abstracts.html and http://www.zurich.ibm.com/pdf/privacy/report3-final.pdf

[72] Privacert Risk Assessment Server (licensed to Privacert, Inc. by L. Sweeney, Carnegie Mellon University). http://privacert.com/assess/index.html

[73] Sweeney L. Patient Privacy Risks in U.S. Supreme Court Case Sorrell v. IMS Health Inc.: Response to Amici Brief of El Emam and Yakowitz. Data Privacy Lab Working Paper 1027-1015B. Cambridge 2011. http://dataprivacylab.org/projects/identifiability/1027.html

seek less privacy-invasive approaches or to augment its approach with traditional remedies (e.g. Fair Information Practices or informed consent), and showed no interest in exploring new promising scientific or societal approaches to privacy protection. *Once data are deemed de-identified under HIPAA, under either the Safe Harbor provision or the Statistician provision, the data can be shared widely for any purpose*.

What is needed is to continue to invest in data privacy research and to establish channels for NCHS, NIST or a professional data privacy body to operationalize research results so that data sharing decisions rely on the latest guidelines and best practices. Doing so will not only improve data sharing practices but will also introduce many other forms of provable privacy protections.

Lack of Fitness for Other forms of Medical Data

De-identification provisions for the HIPAA Privacy Rule were designed narrowly with field-structured, person-specific claims data in mind. This perspective severely limits the ability of the provisions to apply to other forms of research data, even other forms of medical data. For example, clinical notes and letters between physicians are textual documents containing rich references to the lives of patients even when the Safe Harbor provisions are removed. For example, "this first occurred when she danced the lead to Showboat causing her to miss the first month" is the kind of references to employment and lifestyle commonly occurring in clinical notes and physician letters.[74],[75] While often uniquely identifying, it does not require further redaction to be released in accordance with the HIPAA Safe Harbor provision. As another example, the HIPAA Safe Harbor provision requires dates to only reveal the year, and it does not impose any restrictions on transmission time or timestamps. So, consider a clinic that each day transmits a full day of events with time stamps from the previous day; even though the date reports only the year, one can infer the actual month, day and year of the events. Images and genomic information have problems too.

On the other hand, there have been scientific advances in ways to provide aggregate statistics, synthetic data, contingency tables, and other generalized knowledge with guarantees of anonymity e.g., [76] and [77], yet there is no incentive in HIPAA to use these approaches when practical because the HIPAA Safe Harbor provision allows more detailed data to be shared. As scientists develop more innovative remedies, there should be incentives established and distribution channels for NCHS, NIST or a professional

---

[74] Sweeney L. Replacing Personally-Identifying Information in Medical Records, the Scrub System. In: Cimino, JJ, ed. Proceedings, Journal of the American Medical Informatics Association (AMIA). Washington, DC: Hanley & Belfus, Inc, 1996:333-337. http://dataprivacylab.org/projects/scrub/index.html

[75] Clifton C et al. Anonymizing Textual Data and its Impact on Utility. http://projects.cerias.purdue.edu/TextAnon/

[76] Cynthia Dwork. Differential privacy: A survey of results. In Theory and Applications of Models of Computation, TAMC 2008, volume 4978, pages 1–19. Springer, 2008.

[77] Sweeney L. Demonstration of a Privacy-Preserving System that Performs an Unduplicated Accounting of Services across Homeless Programs. Data Privacy Lab Working Paper 902. Pittsburgh 2007, October 2008. http://dataprivacylab.org/projects/homeless/index2.html

data privacy body to operationalize research results so that real-world data sharing decisions rely on the latest guidelines and best practices.

Impact of Commercial Data Sharing on Researchers

HIPAA provisions were crafted from the perspective of governing the data source (e.g., hospital, physician, insurance company) and not the data recipient. Most researchers have historically been data sources, compiling information from observations, surveys and experiments, but increasingly, many researchers are no longer data collectors, but analyzers of data already collected. This fundamental shift places limits on the exposure to HIPAA litigation, criminal, and civil risks that researchers and research organizations may be willing to bear without seeking an alternative research structure that would not have such risk. *Research organizations that primarily rely on corporate data holders may form as a way of opting out of government imposed privacy oversight if HIPAA provisions are heavily imposed.*

Our understanding of ourselves is beginning to be transformed by computational social sciences and by genomics.[78] The reason is personal data: as we move through our lives we leave continuous, multifaceted digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.[79] Researchers who work in these emerging areas typically use data collected elsewhere, by corporations not bound to HIPAA or IRB regulation.

Many companies thrive through selling products and services that are enabled through the acquisition, curation and aggregation of personal data. For example, IMS Health collects personal prescription information from pharmacies and pharmacy benefits programs, and then uses it to sell market information to pharmaceutical companies.[80] Acxiom collects personal information from public records, such as marriage licenses and voter lists, and uses it to provide background checks.[81] Geisinger Health System, a large integrated health system, created a company called MedMining, which licenses its data to promote healthcare research, primarily to major pharmaceutical companies and large biotech companies.[82] Other companies (e.g. Google and Facebook) trade the use of online services for access to personal data.

Research access to commercial data can be unencumbered. For example, when Latanya Sweeney pioneered early research on finding and replacing personal information in

---

[78] Lazer D, Pentland A, Adamic L et al. Computational Social Science. Science. 323(6). Feb 2009. pp.721-722.
[79] Pentland A. Honest Signals: How They Shape Our World, Chapter 7, pp. 85-94, MIT Press, Cambridge, MA. 2008.
[80] IMS Health. IMS Facts at a Glance. As of September 30, 2010, http://www.imshealth.com/
[81] Acxiom. FAQs and EEOC Guidelines. As of September 30, 2010 http://www.acxiom.com/products_and_services/background_screening/faq/Pages/FAQs.aspx
[82] MedMining. Welcome to MedMining. As of September 30, 2010 http://www.medmining.com/

textual clinical notes[83], she visited local area hospitals and left with data the same day, acquiring the data through the business office because her work was seen as a way to protect against possible litigation. In contrast, Peter Szolovits at MIT now reports spending 9 months of ongoing negotiations and delays to get the same kind of data from the same hospitals, impeding his research funded by the National Institutes of Health aimed at helping hospitals share data more freely with guarantees of patient anonymity.

Data Privacy, the field

Data Privacy is the study of risk and utility in data sharing arrangements. The question the discipline of Data Privacy seeks to answer is "For given data sharing arrangements, how can we construct integrated techno-policy systems that optimally minimize risk and maximize utility?" The discipline of Data Privacy may involve weaving traditional policy formations into existing technology, or creating innovative new technologies and policy altogether, thereby making it possible for Data Privacy to transcend the false belief that society must choose between privacy or utility, and instead pioneer new solutions so that society can enjoy both privacy and utility.

Over the past 50 years the study of Data Privacy has grown from the efforts of a handful of statisticians exploring ways to render data anonymous and a handful of policy makers that largely ignored mathematical considerations when designing policies for sharing personal data widely, to an emerging broad cross-disciplinary field that has produced fundamental computational theories of anonymity, has designed algorithms for risk assessment and management, has introduced new policy approaches appropriate for a data rich networked society, and has spun off an industry of privacy technologies.

The de-identification provisions of the HIPAA Privacy Rule do not take advantage of advances in data privacy or the nuances it provides in terms of dealing with *different kinds of data* and finely *matching sensitivity to risk*. There needs to a channel for NCHS, NIST or a professional data privacy body to operationalize research results so that data sharing decisions rely on the latest guidelines and best practices. Doing so will not only improve data sharing practices but will also introduce many other forms of privacy options.

-- next question starts on next page --

[83] Sweeney L. Replacing Personally-Identifying Information in Medical Records, the Scrub System. In: Cimino, JJ, ed. Proceedings, Journal of the American Medical Informatics Association (AMIA). Washington, DC: Hanley & Belfus, Inc, 1996:333-337. http://dataprivacylab.org/projects/scrub/index.html

Question 55. *What mechanism should be used to regularly evaluate and to recommend updates to what is considered de-identified information? Beyond the mere passage of time, should certain types of triggering events such as evolutions in technology or the development of new security risks also be used to demonstrate that it is appropriate to reevaluate what constitutes de-identified information?*

Response:

Some of the Executive Summary responds to this question, so it is reprinted and adapted below. Responses to questions 63, and 64 appear thereafter.

The de-identification provisions of the HIPAA Privacy Rule do not take advantage of scientific advances in data privacy or the knowledge it provides in terms of dealing with different kinds of data and finely matching sensitivity to risk. There needs to a channel for NIST or a professional data privacy body to operationalize research results from data privacy research so that data sharing decisions rely on the latest guidelines, methods, and best practices. Doing so will not only improve data sharing practices but will also introduce many other forms of provable privacy protections so that society may enjoy widespread data sharing with privacy protections.

Data Privacy is the study of risk and utility in data sharing arrangements. The question the discipline of Data Privacy seeks to answer is "For given data sharing arrangements, how can we construct integrated techno-policy systems that optimally minimize risk and maximize utility?" The discipline of Data Privacy may involve weaving traditional policy formations into existing technology, or creating innovative new technologies and policy altogether, thereby making it possible for Data Privacy to transcend the false belief that society must choose between privacy or utility, and instead pioneer new solutions so that society can enjoy both privacy and utility.

Over the past 50 years the study of Data Privacy has grown from the efforts of a handful of statisticians exploring ways to render data anonymous and a handful of policy makers that largely ignored mathematical considerations when designing policies for sharing personal data widely, to an emerging broad cross-disciplinary field that has produced fundamental computational theories of anonymity, has designed algorithms for risk assessment and management, has introduced new policy approaches appropriate for a data rich networked society, and has spun off an industry of privacy technologies.

There have been scientific advances in ways to provide aggregate statistics, synthetic data, contingency tables, and other generalized knowledge with guarantees of anonymity e.g., [84] and [85], as well as discoveries on assessing re-identification risks.[86] As scientists

---

[84] Cynthia Dwork. Differential privacy: A survey of results. In Theory and Applications of Models of Computation, TAMC 2008, volume 4978, pages 1–19. Springer, 2008.

[85] Sweeney L. Demonstration of a Privacy-Preserving System that Performs an Unduplicated Accounting of Services across Homeless Programs. Data Privacy Lab Working Paper 902. Pittsburgh 2007, October 2008. http://dataprivacylab.org/projects/homeless/index2.html

develop more innovative ways to provide proofs of risks and remedies, there should be channels for NIST or a professional data privacy body to operationalize research results so that data sharing decisions rely on the latest guidelines and best practices.

-- next question starts on next page --

[86] Privacert Risk Assessment Server (licensed to Privacert, Inc. by L. Sweeney, Carnegie Mellon University). http://privacert.com/assess/index.html

Question 63. *Given the concerns raised by some that even with the removal of the 18 HIPAA identifiers, reidentification of de-identified datasets is possible, should there be an absolute prohibition against re-identifying deidentified data?*

Response:

Some of the Executive Summary responds to this question, so it is reprinted and adapted below. Response to question 64 appears thereafter.

Understanding re-identification risks exposes threat models, their likelihood of success, and if successful, the extent of adverse impact that could result. Prohibiting research on re-identification would drive re-identification methods further into hidden, commercial activities and deprive the public, the research community and policy makers of knowledge about re-identification risks and potential harms to the public. Understanding the risks to re-identification are important to understanding scientific privacy remedies.

At present, the de-identification provisions of the HIPAA Privacy Rule do not take advantage of advances in data privacy or nuances data privacy solutions may provide in addressing different kinds of data sharing arrangements and in finely matching sensitivity to risk. There needs to be a channel for NIST or a professional data privacy body to operationalize research results so that data sharing decisions, whether through regulation or an IRB, relies on the latest guidelines and best practices. Doing so will not only improve data sharing practices but introduce many other forms of provable privacy protections.

-- next question starts on next page --

Question 64. *For research involving de-identified data, is the proposed prohibition against a researcher reidentifying such data a sufficient protection, or should there in some instances be requirements preventing the researcher from disclosing the deidentified data to, for example, third parties who might not be subject to these rules?*

Response:

Rather than a blanket decision on de-identified data, decisions about whether a data use agreement should be used and the terms of the data use agreement should be nuanced on scientific knowledge of risks and remedies specific to the data sharing arrangement. Work in the field of data privacy has already provided technologies that can help assess actual risks and pose less-risky alternatives for sharing data.

At present, the de-identification provisions of the HIPAA Privacy Rule do not take advantage of advances in data privacy or nuances data privacy solutions may provide in addressing different kinds of data sharing arrangements and in finely matching sensitivity to risk. There needs to be a channel for NIST or a professional data privacy body to operationalize research results so that data sharing decisions, whether through regulation or an IRB, relies on the latest guidelines and best practices. Doing so will not only improve data sharing practices but introduce many other forms of provable privacy protections.