Prepared Statement of Ira Rubinstein

Senior Fellow, Information Law Institute, New York University School of Law

Adjunct Professor, New York University School of Law

National Committee on Vital & Health Statistics; Subcommittee on Privacy, Confidentiality, & Security

Hearing on De-Identification and the Health Insurance Portability and Accountability Act

Tuesday, May 24, 2016

Good morning and thank you for inviting me to present testimony to the committee on the topic of "Policy Interpretations of HIPAA's De-identification Guidance." My name is Ira Rubinstein and I am a Senior Fellow at the Information Law Institute, New York University School of Law, and an Adjunct Professor at New York University School of Law, where I teach courses in privacy law. I recently co-authored a paper entitled "Anonymization and Risk" with Professor Woodrow Hartzog, Associate Professor, Samford University's Cumberland School of Law, which will be published in the next issue of the Washington Law Review. My remarks will reflect the broader analysis of these topics in this article. In particular, I will focus my attention on how our analysis applies to several of the specific questions that were posed to the panelists prior to this hearing.

1. What issues do you see as being the most pressing when you consider de-identification and HIPAA?

The HIPAA de-identification rule presupposes that properly de-identified data sets do not reveal the identity of individuals connected to the data and therefore pose no risk to anyone's privacy. In the past twenty years, however, researchers have shown that individuals can be identified in many different data sets once thought to have been fully protected by means of de-identification. In particular, a trio of well-known cases of re-identification has called into question the validity of the de-identification methods on which the HIPAA Privacy Rule also relies. (The three cases famously involve the public release of de-identified hospitalization records of state employees including then-Massachusetts Governor Weld, twenty million search queries of 650,000 AOL users, and more than 100 million ratings from over 480,000 Netflix customers on nearly 18,000 movie titles.)

This so-called "failure of anonymization" has led many critics to argue that de-identification no longer successfully protects data subjects from potential privacy harms when such data sets are publicly released. They point out that for any public release of data, other data sets containing related data will inevitably be released, allowing an attacker to link data in both sets and re-identify individuals in the first data set. Indeed, the researchers who co-authored the study showing how to de-anonymize the Netflix data set using auxiliary data (in this case, the records of about fifty users of the publicly available Internet Movie Database) made the even stronger claim that *any* attribute can be identifying in combination with others. This is a potentially devastating objection to HIPAA-compliant de-identification methods as well. Of course, de-identification has its defenders, too, who counter that despite the theoretical and demonstrated ability to mount linkage attacks based on auxiliary information, the likelihood of re-identification for most data sets remains minimal and, as a practical matter, most data sets will remain de-identified using HIPAA-approved techniques.

I am neither a statistician nor a computer scientist and therefore I will refrain from commenting on the technical aspects of this debate or on which side has the better arguments. Rather, my expertise is in privacy and security law and policy, and from this perspective, I wish to call attention to two serious concerns with the current HIPAA de-identification rule:

(a) *Crisis of faith*. The possibility of correctly identifying people and attributes from de-identified data sets has sparked a crisis of faith in the validity of de-identification methods. Do these methods still protect data subjects against possible privacy harms associated with revealing sensitive and non-public information as a consequence of linkage attacks? Certainly, there is widespread skepticism about de-identification techniques among some leading privacy scholars and most of the popular press, which in turn undermines the credibility of the exemptions for de-identified data under both the safe harbor and the expert determination standards of the HIPAA Privacy Rule. This is of obvious concern because it not only creates legal and regulatory uncertainty for the scientific research community but may even discourage individuals from contributing data to new research projects. (It also heightens consumer mistrust of e-commerce firms offering their own dubious "guarantees" of anonymization, thereby reinforcing the "privacy is dead" meme.)

(b) *Scientific discord*. One would like to think that the relevant scientific experts would have sorted out their differences and resolved these doubts about de-identification by now but that is not the case. In the paper I co-authored with Professor Hartzog, we found that the community of computer scientists, statisticians, and epidemiologists who write about de-identification and re-identification are deeply divided, not only in how they view the implications of the auxiliary information problem, but in their goals, methods, interests, and measures of success. Indeed, we found that the experts fall into two distinct camps. First, there are those we identified as "pragmatists" based on their familiarity with and everyday use of de-identification methods and the value they place on practical solutions for sharing useful data to advance the public good. Second, there are those we called "formalists" because of their insistence on mathematical rigor in defining privacy, modeling adversaries, and quantifying the probability of re-identification. We found that pragmatists devote a great deal of effort to devising methods for measuring and managing the risk of re-identification for clinical trials and other specific disclosure scenarios. Unlike their formalist adversaries, they consider it difficult to gain access to auxiliary information and consequently give little weight to attacks demonstrating that data subjects are distinguishable and unique but that (mostly) fail to re-identify anyone on an individual basis. Rather, they argue that empirical studies and meta-analyses show that the risk of re-identification in properly de-identified data sets is, in fact, very low.

Formalists, on the other hand, argue that efforts to quantify the efficacy of de-identification "are unscientific and promote a false sense of security by assuming unrealistic, artificially constrained models of what an adversary might do." Unlike the pragmatists, they take very seriously proof-of-concept demonstrations of re-identification, while minimizing the importance of empirical studies showing low rates of re-identification in practice.

This split among the experts is concerning for several reasons. Pragmatists and formalists represent distinctive disciplines with very different histories, questions, methods, and objectives. Accordingly, they have shown little inclination to engage in fruitful dialogue much less to join together and find ways to resolve their differences or place de-identification on firmer foundations that would eliminate or at least reduce the skepticism and uncertainty that currently surrounds it. And this makes it very difficult for policy makers to judge whether the HIPAA de-identification rules should be maintained, reformed, or abandoned.

2. Is the current HIPAA de-identification guidance sufficient? Does it pose challenges to or does it advance the use of data (including aggregation, analysis, dissemination, sharing) in healthcare?

In light of the concerns expressed above, it seems clear that the current HIPAA de-identification guidance is not sufficient. At the very least, it should acknowledge and address the anonymization/de-identification debate and/or consider alternative methods of protecting privacy while preserving research utility.

My co-authored paper suggested that the best way to move beyond the current de-identification impasse would be for Health and Human Services (HHS) and other regulators to incorporate the full gamut of Statistical Disclosure Limitation (SDL) methods and techniques into the Privacy Rule, rather than relying almost exclusively on de-identification techniques. SDL comprises the principles and techniques that researchers have developed for disseminating official statistics and other data for research purposes while protecting the privacy and confidentiality of data subjects. Our paper cites an important article by Satkartar Kinney describing SDL in terms of three major forms of interaction between researchers and personal data: direct access (which covers access to data by qualified investigators who must agree to licensing terms and access data sets securely); dissemination-based access (which includes de-identification), and query-based access (which includes but is not limited to differential privacy. In our view, Kinney's work helps to clarify several contested issues in the current de-identification debate. First, as Kinney points out, the most urgent need today is not for improved de-identification methods alone but also for research that "provides agencies with methods and tools for making sound decisions about SDL." Second, her taxonomy calls attention to the fact that researchers in statistics and computer science pursue very different approaches to confidentiality and privacy and all too often do so in isolation from one another. They might achieve better results by collaborating across methodological divides. Third, the legal scholars who have written most forcefully on this topic tend to evaluate the pros and cons of de-identification in isolation from other SDL methods. We believe that debates focusing exclusively on the merits or demerits of de-identification are incomplete and hence that it is time for HHS to consider broadening the Privacy Rule by incorporating the full spectrum of SDL techniques into its regulatory toolkit.

3. What are the points of confusion or challenges related to HIPAA? What are options for resolving these?

An obvious corollary to shifting from a narrow focus on de-identification to the full spectrum of SDL techniques would be to adopt a risk-based approach that tailors the any given use of SDL and related legal mechanisms to an organization's anticipated privacy risks. If HHS fully embraced a risk-based approach, this would transform the HIPAA Privacy Rule into something more closely resembling the law of data security. Our article argues that the Privacy Rule might achieve better results if it incorporated its current focus on de-identification into a process-based, contextual, and risk tolerant security regime.

- *Process-based*: This implies that organizations engaged in releasing data to internal, trusted, or external recipients would assume responsibility for protecting data subjects against privacy harms by imposing technical restrictions on access, using adequate de-identification procedures, and/or relying on query-based methods, all in combination with legal mechanisms, as appropriate.
- *Contextual*: A risk-based approach to data release policy is inherently contextual, because sound methods for protecting released data sets are always contingent upon the specific scenario of the data release. There are at least seven variables to consider in any given context, many of which have been previously identified in reports by the National Institute of Standards and Technology (NIST) and others. They include data volume, data sensitivity, type of data recipient, data use, data treatment technique, data access controls, and consent and consumer expectations.
- *Tolerant of risk*: The field of data security has long acknowledged that there is no such thing as perfect security. If the Weld, AOL, and Netflix incidents prove anything, it is that perfect anonymization also is a myth. By focusing on process instead of output, data release policy can aim to raise the cost of re-identification and sensitive attribute disclosure to acceptable levels without having to ensure perfect anonymization.

What would a full-fledged, risk-based Privacy Rule look like? The obvious place to turn for guidance is the law of data security as set out by the Federal Trade Commission (FTC). The FTC requires that companies collecting personal information provide *reasonable* data security, defined as having four major components: (1) assessment of data and risk; (2) data minimization; (3) implementation of physical, technical, and administrative safeguards; and (4) development and implementation of a breach response plan.

Our paper proposes that these four tenets of reasonable data security can be modified to establish a general requirement that organizations provide "reasonable data release protections." The tenets of reasonable, process-based, data release protections would look similar to those of data security: (1) assess data to be shared and risk of disclosure; (2) minimize data to be released; (3) implement reasonable de-identification and/or additional data control techniques as appropriate; and (4) develop a monitoring, accountability, and breach response plan.

These requirements would be informed by the nascent industry standards under development by NIST and others, including accepted de-identification and SDL techniques as well as a consideration of the seven risk vectors described above. This approach is context-sensitive and would allow organizations to tailor their obligations according to specific risks and only subject them to liability when they adopt ex ante processes that are inadequate in light of these risks.

4. What is your perspective of oversight for unauthorized re-identification of de-identified data?

Of course, those who engage in unauthorized re-identification are also culpable and it might be worthwhile to supplement contractual or statutory obligations not to engage in re-identification with severe civil (or even criminal) penalties for intentional violations that cause harm. It is important that any such statutory prohibitions also include robust exemptions for security research into de-identification and related topics.

5. What are the current gaps in rules (laws, regulations, self-governance regimes, best practices) across sectors, or with respect to de-identification?

The HIPAA Privacy Rule currently outlines two paths for de-identifying health data sets, the Safe Harbor method and expert determinations. As noted above, HIPAA could move closer to process-based data releases in several different ways. First, the Safe Harbor method could be modified to require technological, organizational, and contractual mechanisms for limiting access to de-identified data sets. Additionally, experts might be asked to certify the adequacy of underlying processes rather than merely assess risk in a given case. Under this approach, companies seeking to be certified as HIPAA compliant would be asked to demonstrate that they have implemented a comprehensive data release program analogous to the comprehensive privacy and security programs articulated in FTC consent orders. This would include performing a threat analysis, identifying mitigating controls, and documenting the methods and results of this analysis (as required by the expert determination method). Although these approaches have certain drawbacks, they might better incentivize robust data release protections and mitigate the inherent difficulty of assessing re-identification and sensitive attribute disclosure risk.

Thank you again for this opportunity to address the committee this morning and I hope my remarks have been helpful.

References

1. Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010)
2. Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 29TH IEEE SYMPOSIUM ON SECURITY & PRIVACY 111.
3. Jane Yakowitz, *Tragedy of the Data Commons*, 25 Harv. J.L. & Tech. 1 (2011)
4. Ann Cavoukian & Khaled El Emam, Info. & Privacy Comm'r of Ont., Dispelling the Myths Surrounding Deidentification: Anonymization Remains a Strong Tool for Protecting Privacy (2011)
5. Arvind Narayanan & Edward W. Felten, NO SILVER BULLET: DE-IDENTIFICATION STILL DOESN'T WORK (2014), http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf
6. Satkartar K. Kinney et al., *Data Confidentiality: The Next Five Years Summary and Guide to Papers*, 1 J. PRIVACY & CONFIDENTIALITY 125 (2009)
7. SIMSON L. GARFINKEL, NAT'L INST. OF STANDARDS & TECH., DE-IDENTIFICATION OF PERSONAL INFORMATION (2015), http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf
8. Press Release, Fed. Trade Comm'n, Commission Statement Marking the FTC's 50th Data Security Settlement (Jan. 31, 2014), http://www.ftc.gov/system/files/documents/cases/140131gmrstatement.pdf
9. Khaled El Emam & Bradley Malin, *Appendix B: Concepts and Methods for De-identifying Clinical Trial Data*, *in* Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk 203, 214 (Inst. of Med. ed., 2015)
10. Deven McGraw, *Building Public Trust in Uses of Health Insurance Portability and Accountability Act*, 20 J. Am. Med. Informatics Ass'n 29, 31 (2013)