Testimony and Statement for the Record of
Simson L. Garfinkel, Ph.D.[1]
Information Technology Laboratory
National Institute for Standards and Technology

Hearing on

"De-Identification and the Health Insurance Portability and Accountability Act (HIPAA)"

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

May 24-25, 2016
Hubert H. Humphrey Building
U.S. Department of Health and Human Services
Washington, D.C.

"Overview and framing of current issues"
9:15 - 10:00 a.m.
May 24, 2016

Executive Summary

As part of its mission of promoting US innovation and industrial competitiveness by advancing measurement science, standards, and technology, NIST is pursuing research in the area of data de-identification. Interest in de-identification extends far beyond healthcare. De-identified data can protect privacy of data subjects, but it can also be re-identified. The identifiability of data in a dataset depends on the difficulty of linking that dataset with data elsewhere in the global data-sphere. Data that appear correctly de-identified today might be identifiable tomorrow based on a future data release. Geographic information is especially difficult to de-identify. The HIPAA Safe Harbor is an approach to de-identifying; while it isn't it does usefully provide data scientists with a bright line rule for de-identifying data. Although most de-identification in healthcare today is based on field suppression and generalization, techniques that employ field swapping and the addition of noise may result in datasets that simultaneously do a better job protecting privacy and can provide researchers with more accurate statistical results.

---

[1] Simson Garfinkel is a computer scientist in the Information Technology Laboratory at the National Institute of Standards and Technology. Prior to joining NIST, Dr. Garfinkel was a tenured associate professor at the Naval Postgraduate School in Monterey, California. Dr. Garfinkel's primary research is in the area of computer security, privacy and digital forensics. He received his doctorate from the Massachusetts Institute of Technology in 2005.

Chair Kloss and members of the subcommittee, thank you for the opportunity to speak today about the de-identification of personal information and the Health Insurance Portability and Accountability Act (HIPAA). My name is Simson Garfinkel, and I am a computer scientist in the Information Technology Laboratory of the National Institute of Standards and Technology (NIST). NIST is a non-regulatory federal agency within the US Department of Commerce. NIST's mission is to promote US innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life. As part of this work, NIST publishes interagency reports that describe technical research of interest. Last October, NIST published NIST Interagency Report 8053, "De-Identification of Personally Identifiable Information,"[2] which covered the current state of de-identification practice. My statement is drawn from that report and supplemented with additional research that has been performed in the interim.

Today there is a significant and growing interest in the practice of de-identification. Many healthcare providers wish to share their vast reserves of patient data to enable research and to improve the quality of the product that they deliver. These kinds of data transfer drive the need for meaningful ways to alter the content of the released data such that patient privacy is protected. Under the current HIPAA Privacy Rule, Protected Health Information (PHI) can be distributed without restriction, provided that the data have been appropriately de-identified—that is, provided that identifying information such as names, addresses, and phone numbers have been removed.

Interest in de-identification extends far beyond healthcare. De-identification lets social scientists share de-identified survey results without the need to involve an Institutional Review Board (IRB). De-identification lets banks and financial institutions share credit card transactions while promising customers that their personal information is protected. De-identification lets websites collect information on their visitors and share this information with advertisers, all the while promising that they will "never share your personal information." Even governments rely on de-identification to let them publish transaction-level records, promoting accountability and transparency, without jeopardizing the privacy of their citizens.

But there is a problem with de-identification. We know that there are de-identified datasets in which some of the records can be *re-identified*—that is, they can be linked back to the original data subject. Sometimes this is because the records were not properly de-identified in the first place. Other times it is because information in the dataset is distinctive in some way that was not realized at first, and this distinctiveness can be used to link the data back to the original identity.

---

[2] Simson Garfinkel, NISTIR 8053, "De-Identification of Personal Information," http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

For example, consider a hypothetical case of a researcher who wants to see if laboratory workers at a university are developing cancer at a disproportionally high rate compared to other university employees. That researcher obtains from the university's insurer a de-identified dataset containing the title, age, and five years' of diagnostic codes for every university employee. It would be sad to learn that a 35-year-old professor was diagnosed with ICD-10 Code C64.1 — malignant kidney cancer—but if there are many 35-year-old professors, that data element might not be individually identifying. On the other hand, if the code is linked with a 69-year-old professor, that person may be uniquely identified. The data would certainly be revealing if the patient's title were UNIVERSITY PRESIDENT instead of PROFESSOR.

One of the challenges that we face is that the same properly de-identified dataset today may not be properly de-identified tomorrow. This is because the identifiability of data depends, in part, on the difficulty of linking that dataset with data elsewhere in the global data-sphere. Considered a de-identified medical record for a patient somewhere in the US with an ICD-10 code of A98.4. That record can lie in wait, like a digital land mine, until some website looking for clicks publishes a list of all people in the people in the US known to have been diagnosed with Ebola. Equipped with this new datum, other information in the de-identified data might now fall into place and single out the patient.

To be fair, each of my examples could be considered a case of improper de-identification. The person who prepared the hypothetical university dataset should have looked at the number of university employees with each title, and removed the title for those jobs that were held by fewer than a critical minimum number of employees—for example, 10. Perhaps all of the university senior executives should have had their titles in the dataset changed to a generic title, such as SENIOR ADMINISTRATOR. This is an example of *generalization*, one of several techniques that's used when de-identifying data. Another technique is called *swapping,* in which attributes are literally swapped between records that are statistically similar. Swapping lets a downstream user of the data perform many statistical operations and generally get the right answer, but it adds uncertainty to the results. In this example, with data swapping there would be no way for a downstream data user to know for sure if the university president has cancer, eczema (L30.9), or the common cold (J00).

Technically, we say that techniques like generalization and swapping decrease the potential for *identity disclosure* in a data release, but they also reduce the *data quality* of the resulting dataset. Other de-identification techniques include adding small random values, called *noise,* to parts of the dataset, and entirely removing, or *suppressing*, specific data columns or rows.

Suppressing columns is the de-identification technique that is easiest to understand. It's also one of the most widely used. In fact, it is the technique that is specified by the HIPAA Privacy Rule's Safe Harbor provision. Under the Rule, a dataset can be considered de-identified if eighteen kinds of identifying information are removed *and* the entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information. These identifiers include *direct*

identifiers such as person's name, address, phone number and social security number. But they also include so-called *indirect-identifiers* such as a person's date of birth, the name of their street, or their city. These are called indirect identifiers because they do not directly identify a person, but they can triangulate on an identity if several are combined.

Geographic information requires special attention, because—just like job titles—some geographic areas are highly identifying, while others aren't identifying at all. The Safe Harbor Rule resolves this disparity by allowing ZIP codes to be included in de-identified data if there are at least 20,000 people living in a ZIP—otherwise only the first three digits of the ZIP code may be included, assuming, once again, that there are at least 20,000 people living within that so-called ZIP3 area.

Other US business sectors have looked at the Safe Harbor Rule and sought to create their own versions. After all, Safe Harbor offers a clear bargain: accept the loss in data quality that comes from removing those 18 data types, and the remaining data *are no longer subject to privacy regulation*. You can give them to researchers, publish them on the Internet, or even sell them to a company that will build statistical models. As long as the data provider doesn't know a specific way that the data can be re-identified, there are no privacy-based limitations on de-identified data at all.

The problem with the Safe Harbor standard is that it isn't perfect. We know that some of the people in a dataset that is de-identified to the Safe Harbor standard *can be re-identified*. One reason is because of statistics—it turns out that in some populations, a few people can identified from just their sex, their ZIP3, and their age in years—all of which are allowed under Safe Harbor.

"In 2010, the Office of the National Coordinator for Health Information Technology (ONC HIT) at the U.S. Department of Health and Human Services conducted a test of the HIPAA Safe Harbor method. As part of the study, researchers were provided with 15,000 hospital admission records belonging to Hispanic individuals from a [particular] hospital system. The data covered the years 2004–2009. Researchers then attempted to match the de-identified records to a commercially available dataset of 30,000 records from InfoUSA, a company that claims to have data on 235 million US consumers. Based on … U.S. Census data, the researchers estimated that the 30,000 commercial records covered approximately 5,000 of the hospital patients. When the experimenters matched using sex, ZIP3, and age, they found 216 unique records in the hospital data, 84 unique records in the InfoUSA data, and 20 records that matched on both sides. The researchers then examined each of these 20 matches and determined the two out of the 20 had the same last name, street address, and phone number. This represents a re-identification rate of 0.013% uniques; the researchers also calculated a re-identification risk of 0.22% uniques

using a more conservative methodology. These rates are not a nationwide average, [however] since they are based on a single ethnic population in a single healthcare system."[3]

This example embodies much of the way that de-identification is performed in the worlds of healthcare, finance, and official statistics. First, direct identifiers are removed from the dataset. Next, indirect identifiers are identified and analyzed to make sure that any specific combination of identifiers ambiguously identifies at least a certain number of individuals. This number is sometimes called "k," a reference to Latanya Sweeney's k-anonymity model.[4]

K-anonymity is not an algorithm that de-identifies data. Instead, it is a framework for measuring that ambiguity of records in a released dataset. Many practitioners will use this number to calculate a *re-identification rate:* if no fewer than 20 records have the same constellation of indirect identifiers, then they say that the re-identification rate is 5%— 1 out of 20 —meaning that an attacker trying to match an identity to one of these ambiguous records in the dataset has a 5% chance of being right.

Another way to measure the effectiveness of a de-identification effort is to give the dataset to a *tiger team* and have the team try to re-identify the data subjects. That's what was done in the ONC study. This approach is effective because the study coordinators know the ground truth—they knew which of the matched records are actually matched individuals, rather than simply being recognized as so-called data "uniques." Many of the well-publicized academic re-identification attacks in recent years have not taken this extra step of verifying the match. Instead, they have asserted that data uniques prove that a dataset can be re-identified. The ONC study shows that attempts to characterize a re-identification rate using matched uniques, rather than verifying against the ground truth, may significantly over-estimate the re-identification rate. In the ONC case, it the re-identification rate would have been over-estimated by a factor of 10.

But another problem with re-identification tests such as this is that they are inherently based on assumptions about what other data are available for the matching. However, as more data—both identified and identifiable—become available, those assumptions may no longer hold. And this is the second problem with the HIPAA Safe Harbor Rule.

As data are more widely used and distributed within our society, and as we learn better how to tease identifiable information out of what was previously thought to be unidentifiable data, we will need de-identification techniques that provide stronger, more measurable privacy

---

[3] NISTIR 8053, De-Identification of Personal Information, Simson Garfinkel, October 2015.
[4] Latanya Sweeney. 2002. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5 (October 2002), 557-570. DOI=http://dx.doi.org/10.1142/S0218488502001648

guarantees than those that are provided by Safe Harbor. That's because there may be other data elements that are not considered identifying by data scientists, but which may be identifiable to a friend, relative, or neighbor. For example, a person who has a constellation of several diseases or accidents over the course of several years may be the only person with that specific combination. This is likely true not just for people who have suffered from rare disorders, but for many members of the population.

In 2012, researchers at Vanderbilt University and the University of Texas—including Bradley Malin, who is on today's first panel—showed that an attacker who obtains 5-to-7 laboratory results from a single patient can use them as a "search key" to find matching records in a de-identified biomedical research database.[5] Medical test results aren't classified as indirect identifiers. But it turns out that each of the individual results in a CBC or CHEM7 test has enough variation that, together, to distinctly identify a person. Of course, you can't identify a person today based on a CHEM7 test taken a year ago. The threat model is much simpler—each medical test is unique, so if all of the tests taken by the same person have the same *pseudonym*, and somewhere else a test results is found that doesn't have the pseudonym but has the patient's name, those test results themselves form the link.

Unfortunately, these databases of medical histories and treatments are precisely the kinds of databases that need to be created and made widely available for projects like the Precision Medicine Initiative to succeed.

One solution, as outlined in the 2012 paper, is to add noise to the non-identifying values before the data are released, just to make sure that they can't be linked with another dataset. Many clinicians and medical researchers are apprehensive about the idea of adding noise—they are afraid that the noise may result in incorrect conclusions being drawn from the data. But the 2012 paper showed that it is possible to add noise in such a way that the clinical meaning of the laboratory tests remains the same.

How much noise should be added? Like the value of *k* or the re-identification rate, that's a policy question, not a technology question. More noise will lower *both* the data quality and the identifiability of the resulting data.

Differential Privacy is a framework that describes a relationship between the amount of noise that's added and the amount of privacy protection that results. One of the intellectually attractive aspects of differential privacy is that its privacy guarantees exists independent of any

---

[5] Reducing patient re-identification risk for laboratory results within research datasets
Ravi V Atreya, Joshua C Smith, Allison B McCoy, Bradley Malin, Randolph A Miller
Journal of the American Medical Informatics Association Jan 2013, 20 (1) 95-101;DOI: 10.1136/amiajnl-2012-001026

past or future data release. Unfortunately, this guarantee can come at a heavy cost to data quality, especially if the noise added in an unsophisticated way.

Differential privacy was developed to support query systems. The basic idea was to allow researchers to perform statistical computations without having the system leak personal information about any individual in the database. These sorts of query systems, sometimes called *trusted enclaves*, are in use today, although rarely with the mathematical formalisms and guarantees that differential privacy provides. One of the field's early discoveries was that there is a quantifiable *privacy budget* that query systems have: only a certain number of questions can be answered, because answering more questions will inevitably compromise privacy.[6]

Several approaches have since been developed to minimize the impact of this privacy budget. For example, instead of answering questions, the entire privacy budget can be spent publishing a new, *synthetic dataset*. Such a dataset can be essentially the original dataset with a few columns dropped and a swapped or altered to protect privacy, or it can be a wholly artificial dataset that is statistically faithful to the original dataset, but there is no one-to-one mapping between any individual in the original data and the data that are made publicly available.

This approach is being used today by the Census Bureau, which has created a synthetic dataset for its Survey of Income and Program Participation. According to the Bureau's website, " The purpose of the SSB is to provide access to linked data that are usually not publically available due to confidentiality concerns."[7] What this means is that for every person in the original dataset there is a corresponding person in the synthetic data set, but only the gender and a link to the individual's first reported marital partner remain unaltered by the synthesis process. All of the rest of the information is changed. Re-identification is not unlikely, it is impossible, since the synthetic people are statistical combinations of the real people.

Synthetic and artificial datasets pose a challenge to researchers and the general public. A synthetic dataset designed to allow research on hospital accidents nationwide might let researchers draw accurate, generalizable conclusions about the impact of training and doctor's work hours on patient outcomes, but make it mathematically impossible to identify specific patients, doctors or hospitals. Such a dataset would be useless for the purpose of accountability or transparency.

Whether or this kind of approach should be used in healthcare will ultimately be a policy question, not a technical one.

Once a de-identification strategy and mechanism are decided upon, they need to be formally evaluated. Do they meet the stated policy goals? Does the software faithfully implement the

---

[6] Cynthia Dwork. Differential privacy: A survey of results. In TAMC, pages 1–19, 2008.
[7] https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html

stated algorithm? Are the statistical privacy guarantees promised by the software actually met? Can the software be used reliably and repeatedly without error? Do the institutions conducting de-identification have the necessary training and procedures in place, to minimize the chance of an improper data release? Will there be ongoing monitoring and auditing to make sure that the assumptions made during the de-identification continue to hold?

Finally, I'd like to briefly mention the de-identification of non-tabular data. This is a significant challenge. Free-format text, photographs, video, and genetic sequences can all contain information that is highly identifiable. There is, nevertheless, a need to be able to legally de-identify these data and share them without the restriction—the same bargain that the HIPAA Privacy Rule provides.

One approach for sharing these kinds of data are technical controls such as data sharing agreements or legal penalties for re-identification attempts. Another approach is synthetic data.

More research is needed to determine if systems could be developed that protect privacy and allow unlimited use of the data. Other research is needed to determine a process that can transform raw data so completely that individuals cannot recognize their own data once they are in the crowd. This would solve the especially difficult problem of preventing re-identification of data elements by close friends and family members. Techniques that prevent self-identification must also maintain the quality of the data. Synthetic data may be the only way to accomplish this goal.

-END-