

Privacy matters: When is personal data truly de-identified?

Jay Cline

July 24, 2009 ([Computerworld](#))

The U.S. Department of Health and Human Services (HHS) is about to rule whether health care entities will need to notify patients if their de-identified data -- patient data that has been stripped of all potential for identifying individuals, which is often used for research and development -- is breached. As it stands now, de-identified data is not subject to the new breach-notification rules imposed by the HITECH privacy provisions of the 2009 American Recovery and Reinvestment Act (ARRA) stimulus package. The debate pits privacy activists on the one side -- who often support notification -- with health care organizations on the other, which say the quality of health care hangs in the balance.

This debate hasn't been getting much attention. That's unfortunate, because the outcome could have broader implications within the U.S. and even around the world. Validating that personal data can be de-identified in a way that still retains commercial and social usefulness could set a precedent for many other privacy-related standards and debates.

The ruling will come amid a flood of medical data-breach notifications in California, the first state to impose this requirement. Since January of this year, 823 medical-data breaches were reported to the state government, of which the state investigated 122 and confirmed 116 as breaches. One of them -- inappropriate staff access into the files of the so-called Octomom -- resulted in the statute's maximum fine of \$250,000. So, the stakes are high on how the question of de-identified data is resolved.

Let me confess my bias. I think the people who came up with the Health Insurance Portability and Accountability Act (HIPAA) de-identification rule should be given a Presidential Medal of Freedom. I'm exaggerating only a little. The rule is one of the most practical innovations in privacy regulations worldwide and arguably has saved lives.

Table 1: HIPAA Direct Identifiers

Any of the data elements below, if associated with health information, makes that information personally identifiable, according to HIPAA. 1.

Names

2. Geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of a ZIP code

3. All elements of dates (except year) for dates directly related to an individual (e.g., date of birth, admission)

4. Telephone numbers

5. Fax numbers

6. E-mail addresses

7. Social Security numbers

8. Medical record numbers

9. Health plan beneficiary numbers

10. Account numbers

11. Certificate/license numbers

12. Vehicle identifiers and serial numbers, including license plate numbers

13. Device identifiers and serial numbers

14. Web universal locators (URLs)

15. IP address numbers

16. Biometric identifiers, including fingerprints and voice prints

17. Full-face photographic images and any comparable images

18. Other unique identifying numbers, characteristics or codes

Source: Minnesota Privacy Consultants

How does de-identification work? According to part 164.514 of the Code of Federal Regulations, a HIPAA-covered entity has two choices if it wants to de-identify patient data and use it for any purpose such as research and development:

1. The "safe harbor" option, where the entity can remove from its data all of the 18 personal identifiers listed by HIPAA (see Table 1); or
2. The "statistical" option, where the entity can hire a statistician to determine which of the 18 identifiers it can retain without creating more than a "very small" risk that the data could be re-identified when publicly released.

HIPAA also provides a third "limited data set" method. Under these criteria, the covered entity can remove 16 of those 18 identifiers, but guard the remaining data with additional security precautions. It can use the dataset for research and development, but the remaining data is still regarded as protected health information (PHI) subject to HIPAA.

No other country has developed a more rigorous or detailed guidance for how to convert personal data covered by privacy regulations into non-personal data (see Table 2). Indeed, clinical-research organizations rarely use the "safe harbor" and "limited data set" options because they must strip out so much information -- particularly dates of service and discharge -- that it makes the remaining data almost worthless from a clinical-research perspective.

What's the benefit to America of this obscure de-identification rule? It enables health care organizations that otherwise wouldn't have been able to use patient data to convert it into a format they can use for a range of other purposes. These other purposes include improving the efficacy of

drugs and medical devices and identifying the optimal places to build new health care facilities.

And I haven't heard of a single case of a de-identified data set being breached by criminals and re-identified. I checked the major running tallies of data breaches -- PrivacyRights.org, Datalossdb.org, the Identity Theft Resource Center and HHS's enforcement reports -- and came up empty.

It's probably because there's far less economic incentive for a criminal to go after medical data instead of credit card information. It's harder to monetize the fact that I know that Judy Smith of Peoria has heart disease -- by filing false claims in her name, for example -- than to have Judy's credit card number and expiration date. If I'm a criminal with advanced data skills and I have a day to spend, I'm going to go after financial data and not health data.

That's why I'm biased in favor of the HIPAA de-identification criteria. I think they advance public health without compromising privacy. But the activists calling for change make some arguments worth considering.

Table 2: De-Identification Outside the U.S.

Privacy laws and guidelines in a handful of jurisdictions outside the U.S. address the topic of anonymization and de-identification of personal data, but none applies the detail and rigor of HIPAA. **Jurisdiction**

Approach

Canada

Organizations can develop their own de-identification criteria. Principle 5, Section 4.5.3 of Canada's federal privacy law, PIPEDA, states: "Personal information that is no longer required to fulfill the identified purposes should be destroyed, erased, or made anonymous. Organizations shall develop guidelines and implement procedures to govern the destruction of personal information."

EU

Organizations can develop their own de-identification criteria. Article 29 Working Party Paper 136, "On the concept of personal data," states in sections 5.3-5.4: "Most anonymised patient records are unique in content, but the chances of re-identification can be extremely small, meaning that the data does not have to be considered personal and subject to the data-protection law. ... Simple decisions on collected data items can greatly enhance the de-identification of data subjects: avoid unnecessary identifiable data (e.g. addresses), properly transform identifying data (e.g. names, unique numbers) into useful but anonymised indices that allow unambiguous grouping and linking of data, transform indirectly identifying items (such as dates) into relative references that cannot be directly linked with observational data."

Privacy laws and guidelines in a handful of jurisdictions outside the U.S. address the topic of anonymization and de-identification of personal data, but none applies the detail and rigor of HIPAA. **Jurisdiction**

Approach

U.K.

Organizations can develop their own de-identification criteria. Chapter 2 of the U.K. Information Commissioner's 2002 guidelines on the "Use and Disclosure of Health Data" states: "If it is never necessary to know the identity of the individuals to whom personal data relates, then the data should be anonymised by removing all personal identifiers. Anonymisation is a permanent process and once anonymised, it will never be possible to link the data to particular individuals. However, permanent anonymisation may not always be acceptable. For instance a researcher may have no need to know the identity of the patients suffering from a particular condition. He or she may, however, need to know that the patient who was diagnosed with the condition on a particular date is the same patient who was diagnosed with a different condition on another date. Pseudonymisation, sometimes described as 'reversible anonymisation' provides a solution. In effect a computer system is used to substitute true patient identifiers with pseudonyms. The true identities are not, however, discarded but retained in a secure part of the computer system allowing the original data to be reconstituted as and when this is required. Typically those making day-to-day uses of pseudonymised data would not have the 'keys' allowing the data to be

Privacy laws and guidelines in a handful of jurisdictions outside the U.S. address the topic of anonymization and de-identification of personal data, but none applies the detail and rigor of HIPAA. **Jurisdiction**

Approach

reconstituted."

Indeed, in *Commonwealth Services Agency v Scottish Information Commissioner*, the House of Lords affirmed that personal data that is "barnardized" -- that is, statistically de-identified -- is no longer subject to the country's data-protection law.

Australia

Organizations can develop their own de-identification criteria. Australia's National Privacy Principles "do not apply to de-identified information or statistical data sets, which would not allow individuals to be identified."

Source: Minnesota Privacy Consultants

They often point to four cases:

- LaTanya Sweeney, an assistant professor of computer science, technology and policy at Carnegie Mellon University, in 2004 paid \$20 for a list of the dates of birth, sex and ZIP codes of voters in Cambridge, Mass. She was able to identify then-governor William Weld's information by linking it to a de-identified set of health-insurance information.
- Philippe Golle of the Palo Alto Research Center in 2006 published a paper that used the 2000 U.S. Census data to estimate that one could use these same data fields -- gender, ZIP code and date of birth -- to uniquely identify 63% of the U.S. population.
- When AOL publicly released a list of about 658,000 anonymous users and the Web searches each made from March to May of 2006, *The New York Times* demonstrated that it was able to identify among the users an unsuspecting widow in Georgia.

- In October of that year, Netflix publicly released a data set containing over 100 million movie ratings by 480,000 anonymous Netflix subscribers for a prize it was running. Arvind Narayanan and Vitaly Shmatikov of the University of Texas at Austin later claimed in a study that this data could be re-identified if certain other information about the movie raters was known.

And last month, two Carnegie Mellon researchers made headlines when they released the results of a test where they were able in fewer than 1,000 attempts to identify all nine digits of the Social Security Numbers of 8.5% of deceased people who were born after 1988.

This academic version of the Black Hat conference -- where hackers try to outdo each other -- has led de-identification purists to gravitate around the so-called "k-anonymity" method of statistical de-identification. Hopefully HHS will back-burner this option, because k-anonymity is to data what chemotherapy is to human tissue: It destroys the good when going after the bad.

According to Columbia University epidemiologist and statistical de-identification expert Daniel Barth-Jones, "The problem with certain de-identification approaches [such as k-anonymity] is that they can badly distort the accuracy of statistical analyses."

"Progress on numerous goals for the government's health IT agenda like quality improvement, patient safety and reducing health disparities could be seriously stunted, or even do more harm than good," he added, "if we aren't conducting our analyses with data that has been de-identified with a rigorous approach for preserving statistical accuracy."

But the growing availability of data on people and improvements in re-identification methodology have nonetheless convinced several privacy advocates that it's time to change the HIPAA de-identification rule. Some have recently submitted public comments on the impending changes to HIPAA. What are they saying?

- The Washington-based Electronic Privacy Information Center, known for advancing many lawsuits in its privacy advocacy, wants to see a

clearer definition of what is a "very small" risk of re-identification and tighter guidelines on the statistical methodologies that can be used to arrive at this determination. "[E]ntities that seek to exempt themselves from the breach reporting requirements may actively seek out 'qualified statisticians' that use methodologies or techniques that minimize their notice obligations," EPIC wrote.

- The Washington-based Center for Democracy and Technology, known for its closer relations with industry, nonetheless wants health care entities to make more use of de-identified data for standard health care operations and not just research. It also wants more control over de-identified data sets. "HHS should consider requiring covered entities to enter into data use agreements with recipients of de-identified data," CDT wrote.
- The New York-based Markle Foundation, which brings together leading health care experts from industry, government and academia, broke from its tradition of offering moderate proposals and is advocating that de-identified data sets be subject to breach notification. "Ideally, we believe that, at a minimum, if there is evidence that de-identified data has been breached in plain text form, individuals whose information was part of the data set should be notified," Markle wrote.

I wonder what social good would be accomplished by sending out breach-notification letters for de-identified or limited data sets that were mishandled. I can just see it:

"Dear Grandma Cline, this is the hospital you just visited. I hope you had a pleasant stay. We regret to inform you that there has been an incident. One of our hospital staff recently lost a USB drive we believe may have contained a set of statistics that included only the date of discharge from the hospital and ZIP code. We use these data for research purposes according to the Authorization for Research form you signed when you were in. We aren't certain whether your date of discharge and ZIP code were included in these statistics, but we were obliged to re-identify everyone who may have been on the USB drive to notify them. If your data was on this USB drive, we estimate there is less than a 1% chance your data could be re-identified by anyone other than researchers at Carnegie Mellon University who may have found the USB drive, which is a small device that is inserted into a computer that you can save data on. We apologize for this oversight."

Hopefully, HHS will heed the lesson of the past six years of data-breach notification and not contribute to the overnotification and needless worry and concern of the American public.

I recently caught up with Judith Beach, chief privacy officer of Durham, N.C.-based Quintiles. When she was of the chief privacy officer at Synergy -- a former subsidiary of Quintiles -- the company commissioned the first-ever statistical de-identification of a HIPAA-covered data set. Since then, Beach has become a national expert on de-identification. What's her take on the situation?

"We all want to know if there has been a serious risk to our personal-health data," she told me. "But if we get notified for all incidents, including those of very low risk, we will become inured to the numerous notifications we are bound to receive."

So what should be done? Three things:

- At the HHS level, take up EPIC's advice and define what a "very small" risk of re-identification is, seriously considering the prevailing industry standard of 1%. But leave it to the statistics profession to determine what are appropriate methodologies to use in different scenarios.
- At the health care entity level, treat de-identified data sets as information that should still be protected like other intellectual property, but not at the level of personal health information.
- At the patient level, we all should stop voluntarily posting our dates of birth and personal health information on social-network sites, or accept the elevated risk to our privacy that this causes.

The path HHS takes will be closely watched by other jurisdictions that have not yet defined their own de-identification parameters. If we arrive in a world where personal data is never truly de-identified, we're going to need a risk-based approach to guide our way forward.

Jay Cline is a former chief privacy officer at a Fortune 500 company and is now president of Minnesota Privacy Consultants. You can reach him at cwprivacy@computerworld.com.