Patient Privacy Risks in U.S. Supreme Court Case *Sorrell v. IMS Health Inc.*

Response to Amici Brief of El Emam and Yakowitz

Latanya Sweeney PhD

Carnegie Mellon University
Harvard University

Abstract

In today's data rich networked society, money and outmoded privacy practices are driving personal data into the vaults of private industry networks, notwithstanding potential harms that can result to data subjects. A classic example is IMS Health ("IMS"), which receives prescription data from pharmacies and sells versions of it to pharmaceutical companies for marketing purposes. IMS relies on what can be the weakest of the HIPAA data sharing provisions, allowing for self-assessed claims of confidentiality. There is no external review of IMS' de-identification process, no public detailed statement describing it, and what is reported about it, exposes known vulnerabilities for re-identifying patients. Once data are deemed de-identified under HIPAA, they can be shared widely for any purpose. A stronger HIPAA provision exists, but presumably IMS does not use it because doing so would thwart linking and the ability to construct longitudinal patient records. During the 8 years of the HIPAA Privacy Rule, society has experienced an explosion in the amount of data collected on individuals, challenging HIPAA's 1990s styled protection. Yet, IMS has expressed desire to adapt or seek less privacy-invasive approaches, which are possible under HIPAA. IMS does not augment its approach with traditional remedies (e.g. Fair Information Practices or informed consent), nor has IMS reported interest in exploring new promising scientific or societal approaches to privacy protection. The Vermont Statute, which prohibits the sharing of prescription records, is an effective privacy guard. Unfortunately, IMS and the Vermont Statute leave society with a false belief that one must choose between a secretive privacy-invasive approach or no data sharing at all, overlooking possible ways for society to reap data sharing benefits with privacy protection.

This paper addresses Respondent's arguments, as supported by an *amici* brief filed by Dr. Khaled El Emam and Jane Yakowitz, which in turn, addressed Petitioner's arguments, as supported by *amici* briefs filed by the Electronic Privacy Information Center, the Electronic Frontier Foundation, the AARP and the National Legislative Association on Prescription Drug Prices, and the Vermont Medical Society, on the limited issue of privacy risks of de-identified patient data that is regulated by the Vermont statute.

Keywords: HIPAA Privacy Rule, identifiability, data privacy, re-identification

SUMMARY

With respect to the case before the Supreme Court of Sorrell ("Petitioner") v. IMS Health, Inc . ("IMS", "Respondent"), this paper addresses Respondent's arguments, as supported by an *amici* brief filed by Dr. Khaled El Emam and Jane Yakowitz ("Respondent *Amici* Brief") [1], which in turn, addressed Petitioner arguments, as supported by *amici* briefs filed by the Electronic Privacy Information Center, the Electronic Frontier Foundation, the AARP and the National Legislative Association on Prescription Drug Prices, and the Vermont Medical Society (collectively, "Petitioner *Amici* Briefs") [2, 3, 4, 5], on the limited issue of privacy risks of de-identified patient data that is regulated by the Vermont statute, 18 VT. STAT. ANN. § 4631 (2010), at issue here ("the Vermont Statute") [6].

In short, the Respondent *Amici* Brief and the Respondent rely heavily on an outdated 1990's setting, ignoring today's data rich networked society, the growing monetization of personal data, and potential and actual patient harms, all of which jointly challenge historical privacy protections. Further, the Respondent *Amici* Brief blurs the different de-identification provisions allowed under HIPAA; an important distinction is necessary because the Respondent relies on a weaker provision, and not the stronger provision heavily discussed in the Respondent *Amici* Brief, making it incorrectly seem that the IMS approach to de-identification adheres to the stronger provision.

In fact, the de-identification approach adopted by IMS does not adequately protect the medical privacy of patients. "Patient Data", as Petitioner's *Amici* Brief states, "includes the prescriber's name and address; the name, dosage, and quantity of the drug prescribed; the date and location at which the prescription was filled; and the patient's age and gender. The only missing element –the patient's actual name – is concealed by a weak cryptographic technique that itself enables re-identification of patients."

Finally, the Respondent *Amici* Brief provides no analytical assessment of the IMS Approach itself, while wrongfully describing published results of actual re-identifications and incorrectly interpreting published results that use emerging standards for assessing privacy risks in data. The Respondent *Amici* Brief also fails to mention that the IMS Approach is not readily disclosed, is self-assessed, and ignores scientific and policy enhancements that could render it less privacy-invasive.

If accepted by the U.S. Supreme Court ("Court"), such arguments would have significant unintended consequences, as they would open the door for widespread sharing of re-identifiable data without encouraging data holders to adopt less privacy-invasive approaches when available. Without addressing what level of scrutiny the Court should apply, this paper encourages the Court to not believe that one must choose between a secretive privacy-invasive approach or no data sharing at all, but instead, encourages the Court to require and incentivize data holders to adopt high standards and to continually improve their approaches as improved knowledge and practices become available so that society can enjoy both data sharing benefits and privacy protection.

By way of background, the Vermont Statute imposes a ban, absent prescriber authorization, on the use and disclosure of Patient Data for the purpose of marketing prescription pharmaceuticals to the prescriber. 18 VT. STAT. ANN. § 4631(d) (2010). Federal privacy law, pursuant to HIPAA, as amended by the Health Information Technology for Economic and Clinical Health Act of 2009, Pub. L. No. 111-005 (2009) ("HITECH"), provides that any patient data contained in Patient Data must be sufficiently deidentified before sharing it beyond the pharmacy that collected the information in the care of the patient. The requirement is that there must be no reasonable basis to believe that the data can identify the patient. 45 C.F.R. § 164.514(a). It is important for the Court to respect this requirement.

RESPONSES

1. The Respondent *Amici* Brief asserts [Page 6 (header)] the "effectiveness of HIPAA de-identification standards already in place," while ignoring today's data rich networked society which provides many new ways to re-identify de-identified data.

In fact, the arguments about the effectiveness of HIPAA de-identification standards advanced in the Respondent *Amici* Brief rely heavily on a 1990's way of thinking about available data --i.e., the re-identification threat is limited to demographic fields of data matched against one other dataset. Dr. Sweeney was first to demonstrate this threat by matching demographics in de-identified medical data to a population register to affix patient names to records in the data. She further showed that 87% of the U.S. population was uniquely identified by {date of birth, gender, ZIP} [7]. In response, the commentary of the HIPAA Privacy Rule explicitly cited her work [8] and sought de-identification practice in which ZIPs and dates are made more general, for example, using age rather than the full date of birth and the first digits of the ZIP rather than all 5 or 9 digits.

Even today, making demographic values (e.g., dates and ZIP) less specific can thwart re-identification threats that reply on demographics alone. The Respondent *Amici* Brief (at 8) describes an empirical test conducted recently by the HHS Office of the National Coordinator for Health Information Technology ("ONC") in which a small team had a set of approximately 15,000 patient records from one ethnic group and attempted to re-identify them by matching them to records in a commercial data repository and other external sources (e.g., InfoUSA) and found only two matches. Note that this test was on data adhering to a higher standard of de-identification than IMS uses, specifically having no dates, no unique patient identifiers (hashed or otherwise), and absent full ZIP codes. As known in 1990, testing these data against a single dataset on demographics alone can be effective. However, as computing power has grown and data storage has become inexpensive, databases can now easily use all or diverse subsets of fields across multiple kinds of datasets for re-identifications, posing challenges not tested in the experiment.

The Respondent *Amici* Brief (at 13) mischaracterizes an even earlier re-identification experiment conducted by Sweeney in 1998 reported in *Southern Illinoisan v. Ill. Dep't of Pub. Health*, 218 Ill. 2d 390 (Ill. 2006) [9]. In *Southern Illinoisan*, an Illinois newspaper requested the Illinois Department of Public Health to release from the Illinois Health and Hazardous Substances Cancer Registry (the "Cancer Registry") copies of documents relating to the incidence of neuroblastoma in Illinois. The plaintiff newspaper requested release of the information in a format showing type of cancer, ZIP code, and date of diagnosis. In 1998, an era predating today's Web and data repositories, Illinois Public Health officials reported that Sweeney accurately provided the names of 20 of the 22 children whose information appeared on the Cancer Registry. The Illinois Supreme Court upheld the Illinois Appellate Court's belief that the method Sweeney used was unique to her education and experience and ordered the method sealed. Sweeney maintains the method would not be surprising if revealed today, 12 years later, and that given today's data rich networked society, a high school student could, in less than an hour, easily re-identify the data using information readily available on the Web.

In 2003, Sweeney demonstrated 2-step re-identifications in which non-demographic fields (e.g., diagnosis and procedures over time) are matched to publicly available medical claims data (e.g. hospital discharge data) to learn patient demographics, and the learned demographics then matched to population registers (e.g., voter lists) to re-identify patients by name [10]. The Respondent *Amici* Brief (at 13) reference a technical report, published in online public repositories by Sweeney (citing Sweeney, Patient Identifiability in Pharmaceutical Marketing Data, Cambridge, Data Privacy Working Paper No. 1015(2011)), and describe the 2-step analysis conducted in 2003 with prescription data as "a complex, multistage" approach rather than the two database instructions described above.

Today, in 2011, much more information is readily available, including for example, location data from mobile phones [11] that can relate identified individuals to specific pharmacy visits and purchases. The computational and informational ability to re-identify data is no longer limited to directly linking on a single dataset or using demographics alone.

This history of documented re-identifications, which includes outside review of actual re-identifications, show that as time has passed, more data has become available about individuals, providing more ways to re-identify de-identified data. Therefore, ways of thinking about de-identification have to evolve also. Unfortunately, IMS and the Respondent *Amici* Brief insist on using 1990's standards narrowly.

2. The Respondent *Amici* Brief ignores the growing monetization of personal data, and potential and actual harms that challenge historical privacy protections.

In fact, Price Waterhouse Coopers predicts that sharing personal health information beyond the direct care of the patient will soon be a two billion dollar market [12]. Examples of companies other than IMS with sales relying on personal data: Acxiom collects personal data from public records, such as marriage licenses, and uses it to provide background checks [13]. Geisinger Health System, a large integrated health system, created a company, MedMining, which licenses its data primarily to major pharmaceutical and biotech companies [14]. Other companies (e.g. Google Health and Microsoft Health Vault) trade the use of online services for access to personal data, including prescription data. According to a local newspaper, the State of Illinois sells personal information to insurance companies, federal and state government agencies, and others, raking in millions of dollars [15].

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 HIPAA, but rarely imposed fines and prosecuted only two criminal cases by 2006 [16]. As of last year 2010, there were 8 HIPAA criminal convictions [17] and a $1 million settlement with Rite-Aid [18]. 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 [19]. There have been allusions to a banker crossing medical information with debtor information at his bank, and if a match results, tweaking creditworthiness accordingly [20]. 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.

3. The Respondent *Amici* Brief asserts [Page 6 (header)] the "effectiveness of HIPAA de-identification standards already in place", and (page 8) that "Patient data that has been properly de-identified pursuant to HIPAA poses very small risks of reidentification."

Respondents take note, as they must, that HHS has doubts about whether the existing de-identification standards are adequate. On page 7, they state:

> HHS also is undertaking a review to confirm whether the specifics of the Safe Harbor Method should be updated to reflect any developments in the marketplace. See HHS.gov, http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/deidentificationagenda.html (last visited Mar. 28, 2011). For all of these reasons, the de-identification standards established by HIPAA properly address changes in technology and the overall information environment, and provide strong protections against re-identification.

The fact that HHS began this review is evidence that the existing de-identification methods may be inadequate. However, the existence of a process that may eventually lead to a change in HIPAA required de-identification methods is not evidence that the existing methods fully protect the privacy interest of patients today. HHS sponsored a conference on the subject in March 2010, and it could propose a change in the de-identification standard at some time in the future. That change would be subject to notice and comment through the usual administrative process. Any actual change made in the rule would likely take effect some months after promulgation of a final rule. The delay from recognition of a problem to the imposition of a new de-identification procedure will be measured in years. In the meantime, the existing insufficient HIPAA de-identification standard is not adequately protecting patient privacy against the possibility that records released lawfully today as de-identified records are being re-identified and used for unlawful or undesirable purposes. It is also the case that any new standard, even if adequate when published, may quickly be overcome by other technical, statistical, and data developments. The threat of re-identification is not solved by a fixed pronouncement by HHS at any given point in time.

4. The Respondent *Amici* Brief asserts (page 9) that "HIPAA also carries strict penalties for noncompliance."

Only after an extended discussion of the HIPAA enforcement provisions do respondents note on page 11 that "HIPAA does not apply to every entity that may ever access or use de-identified data…" The reality is that data de-identified under HIPAA standards may be published and disclosed to any person without restraint. None of the HIPAA penalties or enforcement procedures applies to individuals who take HIPAA de-identified data and seek to re-identify it. Respondents' entire discussion of HIPAA enforcement is not only irrelevant, but it proves the opposite. Any person can take HIPAA de-identified data, re-identify it, and reuse it without any possibility that the United States Government can hold that person accountable under HIPAA. There is no HIPAA remedy against the risk of re-identification because HIPAA de-identified data is not regulated.

5. The Respondent *Amici* Brief asserts the irrelevance of examples about re-identification of supposedly de-identified data that do not involve Patient Data.

From Respondent *Amici* Brief:

> The citations provided highlight general concerns that re-identification of supposedly de-identified data can be accomplished by highly trained experts, virtually all of the citations have no relevance whatsoever to the immediate situation, because they do not address how the use or disclosure of [Patient] Data specifically compromises patient re-identification protections. (page 11)

In fact, the examples are highly relevant because they show that data thought to be de-identified under current standards could be re-identified using methods that were not considered, anticipated, or foreseen by the disclosers of the data. The examples illustrate the general principle that de-identification methods cannot be assumed to work effectively forever. Changes in technology, statistical methods, and data availability all combine to make yesterday's de-identified data identifiable today. Those who may be engaged in re-identifying prescriber data or any HIPAA de-identified data have an incentive to hide their re-identification successes lest the rules be changed so we do not know if current loopholes are being exploited.

6. The Respondent *Amici* Brief analyzes a study by Dr. Sweeney and seeks to dismiss it arguing that "In order for such a re-identification attack to reliably occur and have any reasonable certainty of truly reidentifying individuals, complex statistical modeling requiring highly advanced skills and training would be required…" (page 13)

In fact, it is irrelevant what skills are required to re-identify data. Data cannot be treated as de-identified if it cannot be re-identified by most people. If there are methods that will re-identify data, those methods can be learned, may be automated, and can be spread across the Internet to the far corners of the world. As long as there are incentives to find patients and market to them, the only question for marketers is whether the economic returns justify the costs. Skills to support profitable activities will be found.

Further, those who seek patient data for use in marketing directly to patients may not care if the results of a re-identification process are accurate. It is well known that marketers flood consumers with solicitations. Precision is not essential because a two percent response rate can represent a successful and profitable campaign. The identification of patients with particular health conditions allows marketers to solicit those patients for their life times, and to solicit their families beyond that. A method that finds fifty possible patients for every one real patient may be fully acceptable to marketers as likely to return a profit. For those individuals whose records were accurately re-identified, the consequent loss of privacy can be devastating.

7. The Respondent *Amici* Brief cites approvingly a decision by an Illinois state court that dismissed re-identification efforts by Dr. Sweeney on the grounds that "the methodology she used during her experiment was unique to her education, training and experience not easily duplicated by the general public" (page 17)

In fact, there are several problems here. First, respondents claim that re-identification is not possible and then seek to dismiss those examples where data was, in fact re-identified. Second, whether Dr. Sweeney's methodologies are "unique" is something that the court did not and could not know. Regardless, judicial dicta from an older case says nothing about the capabilities that exist today. Third, since that case was decided, Dr. Sweeney has trained many classes of students in just the techniques that she used. Even if it were true that her skills were unique at an earlier time, it is not true today as classes and academic papers have spread re-identification technology and methodology much farther than before. Fourth, Dr. Sweeney is not the only academic using and teaching skills in this arena. We note that respondent Dr. Khaled El Emam, holds the Canada Research Chair in Electronic Health Information at the University of Ottawa, where he is an Associate Professor in the Faculty of Medicine and the School of Information Technology. Finally, it is irrelevant what skills the general public has with respect to re-identification. An individual whose privacy undermined by actions taken by any one of the numerous trained statistical and other specialists – rather than by a member of the general public – will not care who undertook the re-identification.

8. The Respondent *Amici* Brief argues (on pages 18-19) that the de-identified patient data regulated by the statute in this case is already in use for a variety of purposes.

> Even if Vermont were found to have some form of recognizable interest in regulating the de-identified patient data that is incidentally contained within Patient Data, the Vermont Statute does not advance such interest in any meaningful way. First, the Vermont Statute permits the use and sharing of Patient Data for a variety of purposes, such as: pharmacy reimbursement; prescription drug formulary compliance; patient care management; utilization review by a health care professional, the patient's health insurer, or the agent of either; or health care research.

In fact, with one exception, all of the activities cited by the Respondent *Amici* Brief are carried out by organizations that are HIPAA covered entities or their business associates. These organizations are prohibited by HIPAA from using health data for marketing purposes or for other prohibited activities. It is the sharing of data with organizations that have no obligations under HIPAA that presents the most danger to the privacy of patients. The one exception in Respondent's list is health care research. Health care researchers operate under ethical standards that restrict or prohibit secondary uses of patient data. Health care researchers are also overseen by institutional review boards (IRB). Any health care researcher found misusing patient data through improper re-identification will find it difficult to obtain research grants in the future or to have projects approved by IRBs. Researchers are not the class of data users most likely to exploit patient data for commercial activities such as marketing. What is true of researchers and HIPAA covered entities is not true of others who may have marketplace incentives to exploit de-identified data.

9. The Respondent *Amici* Brief asserts (page 20) that "the use of de-identified data routinely results in vast improvements in the privacy protections for individuals compared to the use of identified data."

In fact, this statement is true, but respondents miss the more important point. The benefits that can result from the use of de-identified data can be fatally and broadly undermined if that data is re-identified. It could take but a single, well-publicized, instance of re-identification to convince policy makers and the public that de-identified data should not be shared for research or other purposes. The threat to the use of de-identified data is greatest where, as with patient data, commercial exploitation of the data by marketers is foreseeable and highly likely. Where patient data is shared under circumstances where there is a significant financial incentive for wholly commercial ventures to find ways to exploit that data for marketing purposes, the risks to the other societally beneficial uses of de-identified data are so great that the legislature may well choose to restrict some more dangerous activities in order to preserve the benefits of less risky ones.

10. The Respondent *Amici* Brief does not properly distinguish between different de-identification provisions allowed under HIPAA, making it incorrectly seem that the IMS approach to de-identification adheres to the stronger provision.

In fact, as The Respondent *Amici* Brief correctly describes, HIPAA has two relevant provisions for data sharing:

> HIPAA establishes a high standard that patient information is "de-identified" only when "there is no reasonable basis to believe that the information can be used to identify an individual." 45 C.F.R. § 164.514 (a). There are two methods to comply with this high standard. The first is more common in the context of the pharmaceutical industry. It requires a formal determination by a qualified statistician who, applying statistical and scientific principles and methods for rendering information not individually identifiable, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information ("The Statistician Provision"). ... The statistician ... must reach a conclusion that the risks of re-identification are "very small" in order for the patient information to be properly "de-identified."
>
> The second method is less common in the context of the industry in this case. It involves the removal of eighteen specified patient identifiers, including but not limited to, patient name, location (other than state or 3-digit ZIP codes with populations greater than 20,000), email address, telephone number, Social Security Number, and the like ("Safe Harbor Provision"). 45 C.F.R. § 164.514(b) (2)(i). Significantly, the eighteenth identifier that must be removed is "any other unique identifying number, characteristic, or code."

IMS relies on the first method above, the Statistician Provision. There are many shortcomings to this provision as currently written. How small is "very small"? What qualifications should a person have? What exactly are the criteria used to make the determination? HIPAA itself provides no answers, and so, any two lay "statisticians" can give wildly different assessments and there are no external guidelines, and no required accountability or publication of the assessment criteria or finding. In fact, IMS self-assesses and has not published details of their approach or analysis.

The second method, the Safe Harbor Provision, is well-defined. It explicitly identifies the fields and values that can and cannot be present in the data. Patient Data does not adhere to the Safe Harbor Provision because it includes full dates, the exact locations of pharmacies, and a unique identifier assigned to each patient. Ironically, all of the examples used in the Respondent *Amici* Brief to demonstrate the strength of HIPAA de-identification rely on the Safe Harbor Provision, even though this is not the provision used by IMS.

11. The de-identification approach adopted by IMS does not adequately protect the medical privacy of patients.

In fact, here are two glaring problems. First, as reported in Petitioner *Amici* Briefs, is the vulnerability of IMS using MD5to compute patient identifiers:

> Verispan [IMS] uses the MD5 Hash Algorithm to conceal the actual identity of patients who receive prescription medications. C.A. App. A99 (trial 22 testimony of Jody Fisher, Vice President of Verispan's Product Management); see also 45 C.F.R. § 164.312(e)(2)(ii) (2010). MD5 was developed by Ron Rivest in 1991. MD5 is a cryptographic "hash function" that creates a fixed length "digest" based on a text input. As such, it is possible to transform a person's name into a unique code and, in theory, not to determine the original name from the resulting code. Ron Rivest, The MD5 Message-Digest Algorithm, RFC 1321 (Apr. 1992) [21]. MD5 is an improved version of MD4 and is similar in design. BRUCE SCHNEIER, APPLIED CRYPTOGRAPHY 436 (2nd ed. 1996).

After a series of vulnerabilities were reported, the Petitioner *Amici* Brief goes on to state:

> The Department of Homeland Security's Computer Emergency Readiness Team concluded that MD5 is "cryptographically broken and unsuitable for further use." Chad Dougherty, Vulnerability Note VU#836068: MD5 Vulnerable to Collision Attacks, United States Computer Emergency Readiness Team (Dec. 31, 2008).[22] Bruce Schneier added that "no one should be using MD5 anymore." Bruce Schneier, Forging SSL Certificates, Schneier on Security (Dec. 31, 2008). [23]

A second problem, notwithstanding a perfect cryptographic means to replace a person's name with a consistent made-up identifier, is the widespread assignment of the same identifiers to the same patients across pharmacies. Anyone having access to the function that replaces patient names with identifiers can generate an index of known names to identifiers and then know which identifiers relate to which names; this is a commonly known as a "dictionary attack." The idea is simple. Run a list of known patient names through the function, recording the identifier that results for each name. Later, when given data containing an identifier, use the list to identify the corresponding patient name. (See [24] for an example.). These problems are in addition to the other concerns raised about the identifiability of the data when matched to other data sources.

Additionally, IMS' approach to de-identification also fails to segment special medical data classes, such as psychiatric and HIV related prescriptions in Patient Data, as described in HITECH, and enables prescriptions belonging to the same patient to be linked over time, thereby constructing longitudinal patient profiles that are typically more identifying than isolated prescriptions.

12. The Respondent *Amici* Brief provides no analytical assessment of the IMS Approach itself, while wrongfully interrupting published results that use emerging standards for assessing privacy risks in data.

As stated earlier, 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 Scientific Standard [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 [7] and reiterated in the Respondent *Amicus* Brief, 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 Respondent *Amici* Brief quoted Michael Stoto talking about the approach in 2003, when he stated "we were not able to ascertain the validity or reliability of the matching methods used." Michael A. Stoto, J. Domingo-Ferrer, L. Franconi, eds., The Identifiability of Pharmaceutical Data: A Test of the Statistical Alternative to HIPAA's Safe Harbor; CD-only annex Privacy in Statistical Databases, Lecture Notes in Computer Science 4302 (2006), see http://explore.georgetown.edu/publications/index.cfm?Action=View&DocumentID=25940 (last visited March 28, 2011). [26]"

In fact, Dr. Stoto's reference in 2003 was when the approach was first launched. A year later, Qunitles became the first to use a version of the approach in real-world practice after careful legal and scientific review [27] and bioterrorism surveillance efforts sought to use the approach more widely. Over the last 6 years, the approach has been used commercially by numerous large insurance and data mining companies and government agencies [28].

In closing, perhaps the most concerning problem is no demonstrated desire by Respondent to improve the privacy protection it provides. IMS has seemingly not sought alternatives to MD5, even though multiparty solutions have been posed as an alternative (e.g., [29]). IMS has not attempted to incorporate Fair Information Practices or informed consent into their operation. And, IMS has shown no reported interest in exploring new promising scientific or societal approaches as they unfold, choosing instead to hold on to an outdated 1990's model.

Acknowledgements

Many thanks to Robert Gellman for contributions and review. No compensation or other consideration by either party was received for this writing.

References

1 El Emam, K. and Yakowitz, J. Respondent Amici Brief. Sorrell v. IMS Health. U.S. Supreme Court. 2011. (Archived at http://dataprivacylab.org/archives/sorrell/1.pdf)

2 Electronic Privacy Information Center. Petitioner Amici Brief. Sorrell v. IMS Health. U.S. Supreme Court. 2011. (Archived at http://dataprivacylab.org/archives/sorrell/2.pdf)

3 Electronic Frontier Foundation. Petitioner Amici Brief. Sorrell v. IMS Health. U.S. Supreme Court. 2011. (Archived at http://dataprivacylab.org/archives/sorrell/3.pdf)

4 AARP and the National Legislative Association on Prescription Drug Prices. Petitioner Amici Brief. Sorrell v. IMS Health. U.S. Supreme Court. 2011. (Archived at http://dataprivacylab.org/archives/sorrell/4.pdf)

5 Vermont Medical Society. Petitioner Amici Brief. Sorrell v. IMS Health. U.S. Supreme Court. 2011. (Archived at http://dataprivacylab.org/archives/sorrell/5.pdf)

6 Vermont statute, 18 VT. STAT. ANN. § 4631 (2010). (Archived at http://dataprivacylab.org/archives/sorrell/6.pdf)

7 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

8 Federal Register, March and December 2000, Health Insurance Privacy and Portability Act (HIPPA)

9 Southern Illinoisan v. Ill. Dep't of Pub. Health, 218 Ill. 2d 390 (Ill. 2006)

10 Sweeney, L. Patient Identifiability in Pharmaceutical Marketing Data. Data Privacy Lab Working Paper 1015. Cambridge 2011. http://dataprivacylab.org/projects/identifiability/pharma1.html

11 Cohen, N. It's Tracking Your Every Move and You May Not Even Know. New York Times. March 26, 2011.http://www.nytimes.com/2011/03/26/business/media/26privacy.html?_r=2&hp

12 PriceWaterhouseCoopers. Transforming healthcare through secondary use of health data. 2009.

13 Acxiom. FAQs and EEOC Guidelines. As of Septemer 30, 2010 http://www.acxiom.com/products_and_services/background_screening/faq/Pages/FAQs.aspx

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

15 Essig, C. Illinois rakes in millions selling personal data. Pantagram.com. April 16, 2010. http://www.pantagraph.com/news/state-and-regional/illinois/article_370b913e-4991-11df-ac1b-001cc4c002e0.html

16 Stein, R. 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

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

18 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/

19 Linowes, D. 1996. “A Research Survey of Privacy in the Workplace,” white paper available from the University of Illinois at Urbana-Champaign.

20 Woodward, B.The Computer-Based Patient Record and Confidentiality. New England Journal of Medicine. 1995; 333:1419-1422

21 Rivest, R. The MD5 Message-Digest Algorithm, RFC 1321 (Apr. 1992)

22 Dougherty, C. Vulnerability Note VU#836068: MD5 Vulnerable to Collision Attacks, United States Computer Emergency Readiness Team (Dec. 31, 2008)

23 Schneier, B. Forging SSL Certificates, Schneier on Security (Dec. 31, 2008)

24 Sweeney, L. Risk Assessments of Personal Identification Technologies for Domestic Violence Homeless Shelters. Carnegie Mellon University, School of Computer Science. Technical Report CMU-ISRI-05-133. Pittsburgh: November 2005. Data Privacy Lab Working Paper 901. Pittsburgh 2005. http://dataprivacylab.org/projects/homeless/index.html

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 Stoto, M. The Identifiability of Pharmaceutical Data: a Test of the Statistical Alternative to HIPAA’s Safe Harbor. In CD-only annex to Domingo-Ferrer J, Franconi L, eds. Privacy in Statistical Databases, Lecture Notes in Computer Science 4302, Springer, 2006.

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. Demonstration of a privacy-preserving system that performs an unduplicated accounting of services across homeless programs. Data Privacy Lab Working Paper 902. October 2007. http://dataprivacylab.org/projects/homeless/index2.html

30 G. Loukides, A. Gkoulalas-Divanis, B. Malin (2010). Anonymization of electronic medical records for validating genome-wide association studies. Proceedings of the National Academy of Sciences, 107 (17) 7898-7903. (doi: 10.1073/pnas.0911686107, http://www.pnas.org/content/107/17/7898.short)