Encouraging the Use of, and Rethinking Protections for De-Identified (and "Anonymized") Health Data

June 2009

This paper advocates for stronger standards for de-identification of health data. Patient data sets have a broad variety of useful applications but must be stringently de-identified in order to maintain patient privacy and overall trust in the health care system. However, technological innovations make it increasingly difficult to protect de-identified data against re-identification. This paper argues in favor of strengthening the current de-identification standard, setting different levels of anonymization for different uses of data, requiring greater accountability for re-identification, and enforcing existing policies that are designed to place limits on the amount of data that can be collected and retained. *

Introduction

The trend towards adoption of health information technology offers substantial benefits not only to individuals in terms of health care quality and efficiency, but also to medical research, public health and other functions that derive value from large sets of health-related data. At the same time, increased electronic flows of health data pose significant risks to privacy. Among the many challenges that will require attention as health IT is promoted over the next few years is how to strip health data of personal identifiers in order to eliminate or reduce privacy concerns, while still retaining information that can be used for research, public health and other purposes.

Under the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, health data that is fully identifiable – data that contains patient names, addresses or other identifiers – is "protected health information" and is subject to restrictions on access, use and disclosure. However, recognizing that aggregate data stripped of identifiers is useful for various purposes, the Privacy Rule establishes two classes of data that are stripped of identifiers and exempts them in whole or part from regulation.

First, the Privacy Rule classifies data as "de-identified" if it has been so stripped of common identifiers that there is no reasonable basis to believe the

* CDT thanks Lygeia Ricciardi, Principal, Clear Voice Consulting, LLC, and Alan Rubel, M.A., J.D., Ph.D., Greenwall Fellow in Bioethics, Health Law and Policy, for their significant contributions to this paper.

information can be re-identified. Under the Privacy Rule, data that qualifies as "de-identified" is not regulated at all. The Rule does not restrict who can acquire it or the purposes for which it can be accessed, used or disclosed.

The Privacy Rule recognizes a second category of data, the "limited data set," that is not fully identifiable. A "limited data set" is stripped of many categories of identifying information but retains information often needed for public health and health research (such as birth dates, dates of treatment and some geographic data). Entities covered by HIPAA may share a limited data set for research, public health and health care operations purposes permitted by the Privacy Rule, so long as all recipients are bound by a data use agreement with the originator of the data.

Although the intentions underlying the Privacy Rule's three-part approach (protected health information, de-identified data, and limited data set) were laudable, the framework has been rendered less satisfactory as a result of technology changes and a growing sophistication in the use of data. At least three challenges arise. First, not all uses of de-identified health data or a limited data set require identical levels of masking. Ideally, a broader spectrum of data "anonymization"[1] options would meet the needs of different contexts and assure that data is accessed or disclosed in the least identifiable form possible for any given purpose.

Second, the Privacy Rule, by permitting use of fully identified data for treatment, payment and "health care operations," provides little incentive for covered entities to use data that is less than fully identifiable for these purposes. Of particular concern is the category of health care operations, which includes some tasks that arguably could be fulfilled with data that is less than fully identifiable. Covered entities are required under the Rule to use the minimum necessary amount of data needed to accomplish health care operations, but CDT is unaware of any circumstances in which this standard has been expressly interpreted to set limits on the identifiability of data used for a particular function.

Third, the de-identification provisions of the Privacy Rule may no longer be as effective as they once were at protecting privacy. Changes in society and technology have made re-identification of health information easier and cheaper than ever before. In addition, the Privacy Rule has never included mechanisms for holding recipients of de-identified data accountable for re-identification.

In this paper we propose several ways to strengthen the Privacy Rule's de-identification standards and to encourage the use of de-identified data through

[1] Throughout this paper, we use the term "anonymized" data to refer to data that is intended to be anonymous to data recipients.

complimentary policies. We also recommend that the Department of Health and Human Services (HHS) consider creating additional data anonymization options (beyond just de-identification and the limited data set), either by regulation or through guidance on how to apply the minimum necessary standard to routine uses of data beyond treatment.[2]

In summary, we recommend that HHS:

- Re-examine the Privacy Rule de-identification provisions (in particular, the safe harbor method for de-identification);
- Strengthen accountability by requiring data use agreements;
- Expand data anonymization options under the Privacy Rule;
- Provide incentives to use less than fully identifiable data for certain purposes;
- Provide support through "Centers of Excellence" in de-identification;
- Require or encourage the use of limited access datasets and other technical solutions;
- Require education and training of staff de-identifying data; and
- Consider increasing public transparency regarding uses of de-identified data.

These recommendations, explained in more detail below, are intended to provide general direction to HHS and other policymakers; each of them will require additional inquiry. The economic stimulus legislation (the American Recovery and Reinvestment Act of 2009) provides at least two vehicles for such inquiry. First, the Secretary of HHS is directed to consult with stakeholders and issue guidance on how to best implement HIPAA de-identification requirements.[3] Second, the Secretary is required to issue guidance on implementation of the HIPAA minimum necessary standard.[4] We hope this paper will help inform those efforts.

The findings and recommendations in this paper are based in part on a one-day workshop held by CDT's Health Privacy Project in September 2008, in which some of the nation's best thinkers on data security and privacy explored issues associated with the de-identification of health data. Participants in the workshop

[2] CDT notes that this was also recommended by the Institute of Medicine's Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule. See Institute of Medicine, Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research (2009) (hereinafter IOM Report), pp 3, 39-40.

[3] ARRA §13424(c).

[4] ARRA §13405(b).

are listed in Appendix A. Except as otherwise noted, the views in this paper are solely those of CDT.

Common Applications of De-Identified Health Data

De-identified health data is used in a variety of ways by a range of public and private entities.[5] Practices involving the use of de-identified health data vary widely. In some instances a single entity or type of entity may use both identifiable and de-identified data in its work. Similar entities pursuing similar goals may take different approaches to handling health data. For example, in the case of public health reporting, some states use de-identified data, while others require that data be linked to patient identifiers.

Among the most widespread applications of de-identified data are the following:

- Quality Improvement – De-identified data is used to assess the results of health care treatments and strengthen the ability of health care organizations to provide better care more efficiently.[6]
- Public Health – De-identified data is used to analyze the causes of disease and to engage in prevention on a community-wide basis. Public health uses include syndromic surveillance, the use of data to detect outbreaks and other health threats before they fully manifest themselves.
- Research – Both clinical and epidemiological research relies on de-identified data (in addition to identifiable data, which is protected by a system of external review boards). A common concern among members of the research community is that the Privacy Rule's de-identification provisions sometimes result in the removal of important detail from data sets.[7]
- Commercial Uses – Many companies use de-identified data to improve their products and support core business operations. For example,

[5] See for example "Draft Secondary Uses of Data and Classification Axes" (2007) by the American Medical Informatics Association (AMIA) Taxonomy Working Group at http://www.amia.org/inside/initiatives/healthdata/2007/taxonomy.pdf. Not all of these uses of data are necessarily limited to data in de-identified form.

[6] According to a national scorecard developed by the Commonwealth Fund, the US health system scored 66 out of a maximum of 100 possible points, painting a picture of "missed opportunities and room for improvement" in healthcare quality and efficiency. See http://content.healthaffairs.org/cgi/content/abstract/hlthaff.25.w457?ijkey=o05rzvque3vQE&keytype=ref&siteid=healthaff.

[7] Remarks by Dr Linda Goodwin of the Duke University School of Nursing at the CDT-sponsored workshop on de-identification of health data, September 26, 2008 (hereinafter "CDT workshop"). Dr Goodwin described the use of de-identified data for research on the prevention of premature births. See also SL Clause, DM Triller, CP Bornhorst, RA Hamilton, and LE Cosler, "Conforming to HIPAA regulations and compilation of research data" in the American Journal of Health-System Pharmacy, Vol 61, Issue 10, 1025-1031 (2004) Available online at http://www.ajhp.org/cgi/content/abstract/61/10/1025.

> pharmaceutical companies use it to characterize population sets, learn which populations are using specific drugs, understand risks to patients, and improve the efficiency of sales.[8]

Although we know that de-identified data is used in these ways, the full extent of use is difficult to determine because de-identified data falls outside the HIPAA Privacy Rule. Thus, there are no limitations on the use of de-identified data or any requirements to track and report sharing or secondary uses. Some institutions carefully weigh the merits of each possible use of de-identified data relative to the risks of re-identification,[9] and many institutions may require data recipients to enter into contractual agreements regarding use of the data. However, there is no way to know how many entities with access to de-identified data take extra precautions.

De-Identification and Limited Data Set Requirements of the HIPAA Privacy Rule

"De-identification" refers to a mechanism by which health data is stripped of potentially identifying information in order to make it extremely difficult to trace any given record or piece of information to an individual person. According to the Privacy Rule, de-identified data is "health information that does not identify an individual and with respect to which there is *no reasonable basis to believe* that the information can be used to identify an individual."[10]

There are two methods whereby data can be de-identified under the Rule: the "statistical" method and the "safe harbor" method.[11] The statistical method requires that someone with "appropriate knowledge of and experience with generally accepted statistical and scientific principles and rendering information not individually identifiable" must determine that the "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."[12] The statistician/expert must document the methods and results of his or her analysis.

The safe harbor method relies on the removal of 18 specific data elements that could uniquely identify an individual, including, for example, name, dates, zip

[8] Remarks of Mark Kohan and Sofia Plotzker, IMS Health, and Stanley W. Crosley, of Eli Lilly and Company at the CDT workshop.

[9] Remarks of Dr Shaun Grannis of the Regenstrief Institute at the CDT workshop. Dr Grannis was describing the protocols of the Indiana Network for Patient Care.

[10] 45 CFR §164.514(a) (emphasis added).

[11] Both terms in quotations are in common usage, but neither is actually named in the HIPAA Privacy Rule.

[12] 45 CFR §164.514(b).

code (except for initial 3 digits in some circumstances), telephone numbers, social security numbers, email addresses or URLs, and license plate numbers. Further, in employing the safe harbor method, a covered entity must not have any "actual knowledge" that the remaining information can be used, alone or in combination with other data, to re-identify patients.

Organizations may assign a code or other means of record identification to allow their de-identified data to be re-identified, presuming they do not share the code and take other precautions to protect it.[13]

According to Dr. Bill Braithwaite, who helped to draft the HIPAA Privacy Rule on behalf of HHS, the safe harbor method of de-identifying data was created as an alternative to the statistical method because most institutions do not have significant statistical expertise. Consequently, there was a need for a "rule of thumb" that could protect privacy while allowing valuable analyses to be carried out.[14] Anecdotally, the safe harbor method is widely used for that reason.

As noted above, the Privacy Rule also includes an alternative to full de-identification—the use of a "limited data set."[15] A limited data set is protected health information that excludes a list of direct identifiers of individuals, similar to but less stringent (specifically with respect to geographic data and dates) than the list of elements to be removed under the de-identification safe harbor method. Unlike fully de-identified data, which can be used for any purpose, a limited data set can be used only for research, public health, or health care operations and only if there is a data use agreement in place between the covered entity that generated the data and the recipient.[16] That is, a limited data set has slightly more information than fully de-identified data, but greater restrictions on how it may be used. (See Appendix B of this paper for a table comparing the de-identification safe harbor standard and the limited data set.)

The limited data set/data use agreement model provides an alternative to an otherwise stark set of choices, but it still may be too restrictive for many public health, research, and health care operations uses because of the amount of identifying data that must be stripped out. Nevertheless, the approach represented in the concept of limited data set – allowing for its use in certain contexts subject to the completion of a data use agreement to bind recipients' use of the data and prevent re-identification and re-disclosure – may be useful to the HHS Secretary in considering how to strengthen the de-identification standard and broadened the use of anonymized or "less identified" data.

---

[13] 45 CFR §164.514(c).

[14] Remarks of Dr. Bill Braithwaite, HIPAA Privacy Rule contributing author at the CDT workshop..

[15] 45 CFR §164.514(e).

[16] Id.

Why a Re-Examination of De-Identification Policy Is Needed

There is no one-size-fits-all de-identification approach appropriate for the universe of health information needs. For example, research on prevention of pre-term births may require the incorporation of calendar dates, while research on drug efficacy may not. Similarly, while syndromic surveillance requires precise geographic data, quality improvement measures may not. However, the Privacy Rule lacks the flexibility needed to adequately meet the diverse needs of data users. The standard for full de-identification often requires stripping out the most useful elements for a given use. The alternative of the limited dataset—in which most, but not all, identifying data is removed—may still provide less information than is needed for a given research, public health, or health operations purpose.

In addition, the fact that under the Privacy Rule de-identified data is entirely free of restrictions, tracking or oversight raises significant concerns. Of most concern to CDT is the lack of protections against, and accountability for, re-identification of de-identified data. Since the Privacy Rule was enacted, changes in technology and data practices have made it significantly cheaper and easier to access, analyze, combine, and re-identify data.[17]

The vast proliferation of digital data points available about an individual makes it easier to establish identity. By one estimate, the average person's medical record, including digital x-rays and scans, contains as many bits of data as 12 million novels—far more than in the past.[18] A statistically unusual pattern, such as a variation in blood pressure, can be used to identify an individual.[19] The advent of genetic testing complicates the picture. One goal of the personalized medicine movement is to ensure that genetic data (in particular, data that is relevant to future diagnosis and treatment) is included in electronic medical

[17] One group of pharmacy researchers tested a set of data de-identified under the safe-harbor method for potential for re-identification. Because the de-identified data contained many unique combination opportunities, the researchers determined that "anticipated [data] recipients, such as physicians, nursing agencies, pharmacies, employers, and insurers…could re-identify their members in the study data set with a moderately high expectation of accuracy." Clause, Steven L., et al, "Conforming to HIPAA Regulations and Compilation of Research Data, American Journal of Health System Pharmacy, (61) (2004), 1025-1031, at 1029. See also Bradley Malin and Latanya Sweeney, "How (Not) to Protect Genomic Data Privacy in a Distributed Network: Using Trail Re-identification to Evaluate and Design Anonymity Protection Systems," Journal of Biomedical Informatics 37 (2004), 179-192; Latanya Sweeney, "Computational disclosure control, a primer on data privacy protection," (2001) available at http://www.swiss.ai.mit.edu.proxy1.library.jhu.edu/classes/6.805/articles/privacy/sweeney-thesis-draft.pdf; Virginia de Wolf et al., "Part II: HIPAA and Disclosure Risk Issues," 28 IRB: Ethics and Human Research 6-11 (2006).

[18] According to IBM as reported by the Wall Street Journal blog in "The Exploding Digital Universe," May 18, 2009 http://blogs.wsj.com/digits/2009/05/18/the-exploding-digital-universe/tab/print/

[19] Remarks by Peter Swire, of the Moritz College of Law of the Ohio State University at the CDT-workshop.

records.[20] Genetic information provides not only a rich (and potentially very sensitive) new source of information about individuals, but is also likely to illuminate information about their relatives.[21]

In addition, members of the public are increasingly sharing health information about themselves in contexts and communities outside of the traditional (and regulated) health environment. Personal health records (PHRs), health blogs, chat rooms, online communities, remote monitoring medical devices, and even social networking sites compound privacy risks. As health IT initiatives create greater ability to link health data across multiple sources, the challenge of ensuring that de-identified data remains anonymous to the data recipient becomes more difficult.

The data explosion goes way beyond health data and genetic information, and includes the huge amounts of data generated in the course of everyday life, much of it only weakly protected by privacy laws or entirely unprotected. According to IDC, a technology market research firm, in 2008 alone the world created 487 billion gigabytes of information, up 73% from 2007.[22] Government agencies at all levels are compiling in digital form data on a wide range of matters, including education, property ownership, residency, and employment.[23] Many of these datasets could in theory be combined and used to link an individual to de-identified health data.

Finally, some have raised concerns about the risk that de-identified data may be used for purposes that may conflict with other public policy goals, even if the data is not ever re-identified. The lack of any tracking or reporting mechanisms for de-identified data makes it difficult to know all of the ways such data is in fact being used, and by whom. [24]

Some Recommendations for Reform

HIPAA de-identification policy needs to be re-examined to ensure that it remains sufficiently rigorous in light of rapidly increasing data availability and is sufficiently protected against re-identification. However, making anonymized data available (and encouraging or requiring its use) for public health, research,

[20] See, for example, Presentation of Brian Munroe, President, Personalized Medicine Coalition, before the 2005 FDA Science Forum,
http://www.personalizedmedicinecoalition.org/programs/munroe_pmc_presentation.pdf .

[21] Remarks of Dr Ken Goodman, of the University of Miami Bioethics Program, at the CDT workshop.

[22] The Wall Street Journal blog in "The Exploding Digital Universe," May 18, 2009
http://blogs.wsj.com/digits/2009/05/18/the-exploding-digital-universe/tab/print/.

[23] Remarks by Dr. Latanya Sweeney, of Carnegie Mellon University, at the CDT workshop.

[24] Remarks by Dr. Mark A. Rothstein of the University of Louisville School of Medicine, at the CDT-workshop.

and day-to-day routine uses like those in health care operations helps to promote information-rich health care and population health while also protecting patient privacy to the maximum extent possible, so long as there are sufficient protections for re-identification. We offer the following specific recommendations to balance the twin interests of flexibility and data protection:

1. Reexamine the HIPAA De-identification Standard

As noted previously, the HIPAA de-identification provisions, which are nearly a decade old, need to be re-examined to ensure that they continue to offer a rigorous methodology for significantly reducing the risk of re-identification. For the most part, this requires a review of the safe harbor method of de-identification, which requires the removal of specific identifiers. The statistical method is designed to be adaptable over time but has the potential to result in less consistent application (and its efficacy depends on the skills of the particular statistician). The standard ideally should be adaptable over time. Any new de-identification guidelines may become obsolete again as technology and the data marketplace evolves. Thus, any new mechanisms to protect de-identified data should be designed to incorporate a regular review process.

De-identification rules also must provide for ease of use for the entities engaged in de-identification of data. De-identification in practice is often much less sophisticated than what might be envisioned at the policy level.[25] Many of the entities that generate health data and bear the responsibility of de-identifying it are not able to handle sophisticated methodologies. They need solutions that allow them to comply with de-identification requirements without a high degree of expertise in-house. Consequently, there will always be a need for a safe harbor-type method of de-identifying data; the key is to strengthen this method and make it durable and scalable over time.

2. Strengthen Accountability through Data Use Agreements

As described previously, the Privacy Rule permits covered entities to use and share de-identified data for any purpose, without any requirement to enter into an agreement defining the terms of data use. As a result, entities receiving de-identified data are under no legal obligation under HIPAA to refrain from re-identifying the data. Given the increased risk of re-identification, the failure of the HIPAA Privacy Rule to include adequate protections against this risk is a significant shortcoming.

[25] Remarks by Dr. Justine Carr, National Committee on Vital and Health Statistics (NCVHS) Work Group on Uses of Health Data, at the CDT-sponsored workshop on de-identification of health data, September 26, 2008.

HHS should consider requiring covered entities to enter into data use agreements with recipients of de-identified data. Such agreements need not rise to the level of business associate agreements, which are needed to protect fully identifiable data. Instead, such contracts can be more limited in scope and similar to those used for limited data sets. Under the current Privacy Rule, a data use agreement between a covered entity and a limited data set recipient must provide that the recipient will not use or share the data for any purposes not covered by the agreement. It must also assure that appropriate safeguards are in place to protect the data, report any aberrations from the terms of the agreement, and agree not to re-identify the data or contact the individuals to whom it pertains.[26] Similar provisions could be required in data use agreements of de-identified data.

In addition, under the current Rule, if the covered entity finds that the limited data set recipient violates any terms of the agreement (assuming the covered entity itself is not able to address the problem), it must stop sharing information with the recipient and report the problem to the HHS Secretary.[27] A covered entity is not in compliance with the Rule if it knew of a pattern of activity or practice of a limited data set recipient that constituted a material breach or violation of the data use agreement and did nothing about it. Similarly, HHS and Congress should consider how to hold entities disclosing or receiving de-identified data accountable when data is inappropriately re-identified.

3. Expand Data Anonymization Options under the Privacy Rule

Different levels of data protections are appropriate in different contexts. Providing only two options for anonymity may limit the value that can be derived from data, leaving researchers and others seeking aggregate data with few alternatives beyond use the of fully identified data. HHS should consider developing additional data set options that can be used for a broader range of research, public health, and operations purposes, and that are appropriately protected against re-identification.

4. Create Incentives to Use Less-Than-Fully-Identified Data

As noted above, the HIPAA Privacy Rule provides little to no incentive for covered entities to use data that has been stripped of some patient identifying information for routine purposes such as health care operations because entities are permitted to use fully identifiable data to meet their needs. The limited data set can be used for this purpose, but it is not clear if covered entities take the

[26] 45 C.F.R. §164.514(e)(4)(ii).

[27] 45 C.F.R. §164.514(e)(4)(iii).

additional step of limiting data identifiability – and entering into data use agreements when the information is shared with outside parties – when doing so is not required. Yet it appears that many health care operations functions could be performed with data that is not fully identified. Use of the least identifiable data should always be encouraged, even where the data access and use is strictly internal.[28]

The economic stimulus legislation requires the Secretary to issue guidance (no later than August 17, 2010) on the Privacy Rule's minimum necessary standard.[29] In developing this guidance, the Secretary should consider whether fully identifiable patient data is needed to accomplish all the activities currently included in health care operations.[30] For example, today covered entities may use fully identifiable data for quality assessment and improvement activities, peer review of health professionals, accreditation or credentialing, performing audits, and business planning. For each of these activities, covered entities need access to data about the care that was provided, but in most cases they do not need information that is identified to a particular patient.

At the same time, the rules governing data that has been stripped of some patient identifiers may not need to be as stringent as what is afforded to fully identifiable health information. For example, disclosure of a limited data set requires a data use agreement, but recipients are not required to comply with every obligation of the Privacy Rule. In developing guidance and considering what protections to apply to data that is not fully identifiable, the Secretary should consider the limited data set model. Ideally, the degree of protection for the data should increase with the degree of identifiability. We recognize that drafting specific rules to accomplish such a sliding scale of protections will be a challenge, given that the policies will still need to be flexible enough to meet

---

[28] Hospitals are often the largest employers in small towns, which makes protecting patient privacy critical even for internal uses of health information. See, for example, Testimony of Claude Earl Fox, M.D., Administrator, Health Resources Services Administration, July 14, 1999, http://www.hhs.gov/asl/testify/t990714c.html.

[29] ARRA §13405(b)(1).

[30] Health care operations include: (1) Conducting quality assessment and improvement activities, population-based activities relating to improving health or reducing health care costs, and case management and care coordination; (2) Reviewing the competence or qualifications of health care professionals, evaluating provider and health plan performance, training health care and non-health care professionals, accreditation, certification, licensing, or credentialing activities; (3) Underwriting and other activities relating to the creation, renewal, or replacement of a contract of health insurance or health benefits, and ceding, securing, or placing a contract for reinsurance of risk relating to health care claims; (4) Conducting or arranging for medical review, legal, and auditing services, including fraud and abuse detection and compliance programs; (5) Business Planning and development, such as conducting cost-management and planning analyses related to managing and operating the entity; and (6) Business management and general administrative activities, including those related implementing and complying with the Privacy Rule and other Administrative Simplification Rules, customer service, resolution of internal grievances, sale or transfer of assets, creating de-identified health information or a limited data set, and fundraising for the benefit of the covered entity. 45 C.F.R. §164.501.

diverse data needs. At a minimum, protections to ensure data is not inappropriately re-identified are critical and must be part of any guidance issued by the Secretary.

Until the Secretary's guidance on minimum necessary is issued, the economic stimulus legislation directs covered entities to use the limited data set when it is possible to do so and still accomplish the purposes for which the data is being accessed, used or disclosed.[31] CDT does not believe this requires entities to always use a limited data set to meet the minimum necessary standard, as the language clearly permits the use of more fully identifiable data where it is needed to accomplish a specific purpose. Nevertheless, covered entities should be encouraged to use limited data sets for health care operations activities wherever such a data set could accomplish the needs for accessing or disclosing the data.

5. Provide Support through "Centers of Excellence"

Given that many HIPAA covered entities do not have the in-house expertise to de-identify data using sophisticated methodologies, HHS should consider designating certain organizations or institutions "centers of excellence" with respect to data de-identification. Covered entities seeking to release de-identified data could be required to consult with these entities to gain the necessary expertise, or can outsource the work of de-identification to such centers. As an alternative, HHS could consider providing incentives for covered entities to rely on the centers for assistance in de-identification rather than simply de-identifying data using the safe harbor method, which even if re-assessed by HHS on a regular basis, will likely always have less statistical rigor. The centers could be independent, licensed non-profits that would oversee the uses of de-identified data, and help to determine what level and methodology of de-identification is appropriate in particular circumstances. They could help to ensure privacy, provide oversight, establish best practices,[32] build stakeholder support, and increase public transparency.[33] As an alternative to establishing independent entities, existing research institutions and major academic medical or technology centers could also apply to be designated as "centers of

---

[31] ARRA §13405(b)(2).

[32] Many private sector companies and organizations do an exemplary job of handling data, not necessarily because of any legal obligation, but because they view it as a business imperative. These Centers could be a mechanism for gathering and disseminating private sector best practices.

[33] These are similar to some of the goals articulated in the AHRQ Request for Information on Data Stewardship Entities released in June of 2007. Federal Register: June 4, 2007 (Volume 72, Number 106), 30803-30805. However, CDT does not believe it is necessary to create a new, single national entity to accomplish these goals. In response to that RFI, CDT's Health Privacy Project endorsed comments submitted by the Markle Foundation's Connecting for Health Collaborative articulating the essential qualities of a governance structure for electronic health information exchange. See http://www.connectingforhealth.org/resources/cfh_ahrq_aqa_rfi_073007.pdf .

excellence."

Any such process created by HHS should include a mechanism for holding such centers accountable for persistently adhering to the criteria required for designation as a center. In developing this process, HHS should also consider partnering with the National Institute for Standards and Technology (NIST), which has significant expertise on data anonymization techniques.

6. Require or Encourage the Use of Limited Access Datasets and Other Technical Solutions

Policies alone are not sufficient to protect privacy. Technical solutions are not a substitute for strong privacy rules but when appropriately applied can play an important role in enforcing policy goals. Relevant in this case are both the particular attributes of a database or program and, at a more general level, the design of an entire technical infrastructure. HHS should consider requiring, or at least encouraging, the use of innovative technical solutions to protect data.

One promising approach is the use of limited access datasets. In common practice today, researchers or others are provided with direct access to data (de-identified or not) and can run queries against it, subject to any applicable research rules (HIPAA with respect to data obtained from covered entities, and the federal "Common Rule" in the case of federally funded research conducted by non-HIPAA covered entities).[34] In the case of a limited access dataset, however, researchers are not given access to the entire data set. Instead, data holders provide aggregate data in response to specific questions as they are posed. Information that is not essential to a particular inquiry, including patient identifiers, is never shared.[35] Thus, for example, rather than allowing a query for exact calendar dates associated with the start and end of a course of medication, a researcher could instead limit queries to the overall length of that course or provide query results only in the least identifiable form (e.g., length of the course of medication rather than exact dates). Similarly, a database or network can return query results with the age of a patient, rather than his or her precise birth date.[36] These measures make it much more difficult to associate data with a particular individual. Examples of limited access data sets that have been made available to researchers are CARDIA, a longitudinal study evaluating the development of cardiac disease in adults funded by the National

[34] For a summary and comparison of the Privacy Rule's research provisions, and the federal Common Rule, see the Institute of Medicine's recent report on research and the privacy of health information, supra note 2.

[35] Remarks of Dr. Cynthia Dwork, of Microsoft Research, at the CDT workshop.

[36] Remarks of Dr. Bill Braithwaite.

Heart Lung and Blood Institute[37] and studies funded by the National Institutes of Mental Health.[38]

In addition, data holders could use tools to help quantify the likelihood (as a percent value) that a given data set can be re-identified so that risk can more easily be weighed against potential benefit. Risk assessment tools such as those developed by the Data Privacy Lab at Carnegie Mellon University can identify data in a particular dataset that is vulnerable to known re-identification inference strategies.[39] Data holders can thus strengthen protections, for example, by aggregating, substituting, or removing data that is useful for known re-identification strategies.[40]

In addition to specific tools and technical protocols, it is critical to underscore the importance of an overall decentralized architecture for maintaining health data, a point that has been repeatedly emphasized in the context of protecting the privacy of health information by the Markle Foundation.[41] The underlying idea is that, rather than constructing one or a few comprehensive databases that would result in great harm to many individuals if they were breached, it is preferable to have data remain where it is originally generated (such as in the physician's office or in a hospital) and pulled together only in response to particular queries or to accomplish a particular health care purpose.

Some have suggested creating or designating specific research databases to facilitate the conduct of research, subject to strong privacy and transparency rules. For example, under Ontario's Personal Health Information Protection Act (PHIPA), health entities may disclose identifiable health data without consent to "prescribed persons or entities" that are prescribed by legislation, including registries maintained for the purpose of improving health care or that relate to organ or tissue donation. Prescribed persons or entities must have in place practices, policies and procedures to protect individual privacy, which are reviewed and approved by the Ontario Information and Privacy Commissioner every three years and must be made transparent to the public.[42] Once personal health information is held by a prescribed entity, that entity may use and disclose information for research purposes. Such research must be approved by a Research Ethics Board if it is in identifiable form, but such

---

[37] http://www.cardia.dopm.uab.edu/lad_use_of_dataset.htm.

[38] http://www.nimh.nih.gov/health/trials/datasets/nimh-procedures-for-requesting-data-sets.shtml.

[39] See http://www.privacert.com for more information.

[40] Remarks of Dr. Latanya Sweeney. See also Sweeney, "Weaving Technology and Policy Together to Maintain Confidentiality," Journal of Law, Medicine & Ethics, 25 (1997): 98-110.

[41] See for example the following frequently asked questions on the Markle website: http://www.connectingforhealth.org/aboutus/faqs.html.

[42] Id.

approval is not required if it is released in de-identified or aggregate form.[43] Currently there are five registries designated as "prescribed persons" under PHIPA.

There are aspects of PHIPA's "prescribed entity" approach that are similar to the above "centers of excellence" and limited data set recommendations. However, CDT has significant concerns about creating additional centralized databases for research purposes, given the enhanced privacy risks associated with such centralized models and significant questions about whether such an approach is feasible in the long term.[44] Conducting research across existing databases, which allows data remain in the place from which it originates, is the most efficient and effective way to meet the needs of our complex health system while protecting privacy and security.

7. Require Education and Training

Any staff involved in de-identifying health data or working with health data that has been de-identified should participate in basic training about how best to protect privacy and security through organizational and technical means. Also essential, of course, are basic physical safeguards, such as locking doors to block access to computers. Basic training, perhaps supported by the "Centers of Excellence," would help to minimize the likelihood of breaches and other misuses of data.

8. Increase Transparency for Uses of De-Identified Data

As previously described, data that has been de-identified according to the Privacy Rule's provisions is free from use restrictions, as long as it is not re-identified. When data has been de-identified and sufficiently protected against re-identification, it does not raise a privacy risk to individuals.[45] However, beyond the privacy issue, and as noted above, some have expressed other policy concerns about the ways that de-identified data is currently being used. To address this issue, policymakers could encourage or require greater public transparency about how data (including de-identified data) is used. Such transparency could contribute to the development of guidelines for regarding data use.

---

43 Id.

44 See, for example, http://www.connectingforhealth.org/resources/cfh_ahrq_aqa_rfi_073007.pdf , page 13 (summarizing concerns about facilitating quality measurement through a national centralized data repository).

45 CDT recently argued this position in an amicus brief filed with the Supreme Court. See http://www.scotusblog.com/wp/wp-content/uploads/2009/04/08-1202.

Conclusion

The expectation of the HIPAA Privacy Rule authors was that the Rule itself (or at least guidance issued to interpret it) would continue to evolve to keep pace with changes in technology and practice.[46] Up until this year, that has not happened. However, the newly enacted economic stimulus legislation requires HHS to make changes to the Rule in a number of areas, and to conduct studies or issue guidance in others. Of particular relevance for this paper is the requirement that HHS re-examine the de-identification standard and issue guidance on compliance with the minimum necessary standard. Both undertakings provide HHS with opportunities to increase privacy protections for patients by expanding the options for use of data that is less than fully identifiable for a range of purposes and to ensure that the de-identification standard remains robust as re-identification becomes easier.

This paper is not an attempt to provide definitive or comprehensive direction for changing de-identification policy, but it does provide some recommendations for promising approaches. Additional research and inquiry in this area will be needed before the ideas laid out in this paper are ready for implementation. This paper should serve as the beginning and not the end of a very important public dialogue.

Developing better practices for the use of aggregated data is important, not only because of its relevance to health care, but because solutions for protecting privacy while benefitting from multiple uses of data are also needed in other sectors, including finance. Health information is often at the leading edge of privacy debates, and solutions found in a health context may be applied much more broadly.[47]

FOR MORE INFORMATION

Please contact: Deven McGraw, Director, CDT Health Privacy Project, (202) 637-9800 x 119, deven@cdt.org

[46] Remarks of Dr. Bill Braithwaite.

[47] Remarks of Peter Swire.

APPENDIX A: September 2008 Workshop on De-Identification, Sponsored by CDT's Health Privacy Project

The following individuals made presentations at the workshop:

- Mark Kohan and Sofia Plotzker, IMS Health
- Bill Braithwaite, MD, PhD – Chief Medical Officer of Anakam, Inc. and HIPAA contributing author
- Justine Carr, MD – Senior Vice President for Quality, Patient Safety, Compliance and Medical Affairs, Caritas Christi Health Care System; Co-Vice Chair, NCVHS Work Group on Uses of Health Data.
- Stanley W. Crosley, JD – Chief Privacy Officer, Eli Lilly and Company; Member of the International Pharmaceutical Privacy Consortium
- Cynthia Dwork, PhD – Principal Researcher, Microsoft Research
- Kenneth W. Goodman, PhD - Professor and Director, University of Miami Bioethics Program; Director, Project HealthDesign Ethical, Legal and Social Issues (ELSI) unit
- Linda Goodwin, RN, PhD – Informatics Program Director, Duke University School of Nursing
- Shaun Grannis, MD, MS – Medical Informatics Researcher at the Regenstrief Institute, Inc. and Assistant Professor of Family Medicine at Indiana University School of Medicine
- Mark A. Rothstein, JD – Herbert F. Boehl Chair of Law and Medicine and Director, Institute for Bioethics, Health Policy and Law, University of Louisville School of Medicine
- Latanya Sweeney, PhD – Associate Professor of Computer Science, Technology and Policy and Director of the Data Privacy Lab, Carnegie Mellon University
- Peter Swire, JD – (Workshop Moderator) Professor of Law at the Moritz College of Law of the Ohio State University, Senior Fellow at the Center for American Progress, and Policy Fellow at CDT

APPENDIX B: Comparison: De-Identification (Safe Harbor) & Limited Data Set

Type of Data	De-Identification	Limited Data Set
Names	Names	Names
Address	All geographic subdivisions smaller than a state, including address & zip (except for initial 3 digits in certain circumstances)	Postal address information, other than town or city, state, and zip code
Dates	All elements of dates directly related to an individual (except for years); special rules with respect to ages of 89 and over.	N/A
Telephone Numbers	Telephone Numbers	Telephone Numbers
Fax Numbers	Fax Numbers	Fax Numbers
E-Mail Addresses	E-Mail Addresses	E-Mail Addresses
Social Security Numbers	Social Security Numbers	Social Security Numbers
Medical Record Numbers	Medical Record Numbers	Medical Record Numbers
Health Plan Numbers	Health Plan Numbers	Health Plan Numbers
Account Numbers	Account Numbers	Account Numbers
Certificate/License Numbers	Certificate/License Numbers	Certificate/License Numbers
Vehicle identifiers & serial numbers (including license plate numbers)	Vehicle identifiers & serial numbers (including license plate numbers)	Vehicle identifiers & serial numbers (including license plate numbers)
Device Identifiers & serial	Device Identifiers & serial	Device Identifiers & serial

Device Identifiers & serial numbers	Device Identifiers & serial numbers	Device Identifiers & serial numbers
Web Universal Resource Locators (URLs)	Web Universal Resource Locators (URLs)	Web Universal Resource Locators (URLs)
Internet Protocol (IP) Address Numbers	Internet Protocol (IP) Address Numbers	Internet Protocol (IP) Address Numbers
Biometric Identifiers, including finger and voice prints	Biometric Identifiers, including finger and voice prints	Biometric Identifiers, including finger and voice prints
Full Face Photographic Images and any Comparable Images	Full Face Photographic Images and any Comparable Images	Full Face Photographic Images and any Comparable Images
Other data	Any other unique identifying number, characteristic, or code, except codes permitted for re-identification	N/A
Standard/Rules for Use	**De-Identification**	**Limited Data Set**
Knowledge of re-identification possibilities	Information is not de-identified if the covered entity has actual knowledge that the information could be used alone or in combination with other information to identify an individual who is the subject of the information.	N/A
Limitation on Uses	N/A	Can be used by a covered entity only for research, public health, or health care operations.
Data Use Agreement Required	No	Yes