OPEN ACCESS

Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data

Deven McGraw

Correspondence to
Deven McGraw, Center for Democracy & Technology, 1634 I Street, NW Suite 1100, Washington, DC 20006, USA; deven@cdt.org

Received 26 March 2012
Accepted 31 May 2012
Published Online First 26 June 2012

ABSTRACT

Objectives The aim of this paper is to summarize concerns with the de-identification standard and methodologies established under the Health Insurance Portability and Accountability Act (HIPAA) regulations, and report some potential policies to address those concerns that were discussed at a recent workshop attended by industry, consumer, academic and research stakeholders.

Target audience The target audience includes researchers, industry stakeholders, policy makers and consumer advocates concerned about preserving the ability to use HIPAA de-identified data for a range of important secondary uses.

Scope HIPAA sets forth methodologies for de-identifying health data; once such data are de-identified, they are no longer subject to HIPAA regulations and can be used for any purpose. Concerns have been raised about the sufficiency of HIPAA de-identification methodologies, the lack of legal accountability for unauthorized re-identification of de-identified data, and insufficient public transparency about de-identified data uses. Although there is little published evidence of the re-identification of properly de-identified datasets, such concerns appear to be increasing. This article discusses policy proposals intended to address de-identification concerns while maintaining de-identification as an effective tool for protecting privacy and preserving the ability to leverage health data for secondary purposes.

INTRODUCTION

Health information collected initially for treatment or payment purposes by healthcare providers and health insurers has high value for other important, 'secondary' purposes, including quality improvement, medical research, public health, and business analytics.[1 2] The ability to access this information from most healthcare providers and health plans is governed by the Health Insurance Portability and Accountability Act (HIPAA) privacy regulations (the privacy rule).[3] But the privacy rule sets standards only for identifiable health information; information that qualifies as 'de-identified' under the privacy rule is not subject to HIPAA regulations.[4] Other federal and state health privacy rules typically also apply only to identifiable information. Consequently, de-identified data are in high demand for a broad range of secondary purposes.

The HIPAA de-identification standards have been controversial since their inception in 2000, and those concerns have increased in the past few years. Such concerns fall into three categories: (1) sufficiency of de-identification methodologies; (2) lack of accountability for unauthorized or inappropriate re-identification; and (3) disapproval of certain uses of de-identified data.

The American Recovery and Reinvestment Act of 2009 requires the Department of Health and Human Services (HHS) to issue a report on the HIPAA de-identification standard.[5] In response, HHS held 2 days of meetings on de-identification in March 2010,[6] but the report had not yet been issued when this article was submitted for publication. Questions continue to linger about the protective value of HIPAA de-identification, while demands for these data increase. In 2011 the USA launched a major federal incentive programme designed to increase the use of electronic medical records by healthcare providers. One goal of the programme is to enhance the quality and efficiency of the healthcare system, which will require greater access to health information for analytics purposes. Failure to address concerns about the de-identification standard effectively could hamper efforts to leverage health information more robustly for health system improvements.

The Center for Democracy & Technology (CDT) began exploring concerns about HIPAA de-identification back in 2009[7] and gathered approximately 50 academic, industry and consumer stakeholders together at a small, non-public October 2011 workshop to vet policy ideas for addressing these concerns. This paper reports on this workshop and explores some of the promising policy proposals in more detail.

Concerns have also been raised about the use of personal information by commercial enterprises operating in the internet and mobile space.[8] Such entities are typically not covered by HIPAA and are not required to comply with its de-identification standards. Discussions of anonymization in those contexts are not covered in this paper, and mentions of 'de-identified' data herein refer to information de-identified per HIPAA.

CONCERNS ABOUT HIPAA DE-IDENTIFICATION: FROM 2000 TO THE PRESENT

The 1996 HIPAA statute authorized HHS to develop rules to protect 'individually identifiable health information' accessed, used or disclosed by 'covered entities' (most healthcare providers, health plans, and health clearinghouses) and their contractors or business associates.[9 10] Under HIPAA, information is identifiable if it identifies the individual or there is 'reasonable basis to believe that the information can be used to identify the individual'.[11] Consequently, the privacy rule defines de-identified data as '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'.[12] Once data qualify as de-identified, they are no longer regulated by HIPAA and can be used for any purpose, without restriction.

In summary, the privacy rule established two methodologies for achieving de-identification: the 'statistical' (or expert) method, requiring a qualified statistician to attest that the data raise very low risk of re-identification, and the 'safe harbor' method, which requires the removal of 18 types of identifiers. Both methodologies were included in the original privacy rule, finalized in December 2000,[13] and have largely remained the same for more than a decade (see box 1).

Concerns about the de-identification standard were raised during its initial development, and they are remarkably similar to concerns that continue to be expressed now. After the standard was initially proposed, HHS received comments stating that 'no record of information about an individual can be truly de-identified' and 'all such information should be treated and protected as identifiable because more and more information about individuals is being made available to the public… that would enable someone to match and identify records that otherwise appear to be not identifiable'.[13] In response, HHS expressly acknowledged the inability to de-identify to a level of zero risk but noted that the statutory standard envisions 'a reasonable balance between the risk of identification and usefulness of the information'.[13]

HHS further declined to deem de-identification status only to information raising zero risk of re-identification because this would 'preclude many laudable and valuable uses'… while imposing 'too great a burden on less sophisticated covered entities to be justified by the small decrease in the already small risk of identification'.[13] Instead, HHS made certain that the easiest path to de-identification, the safe harbor standard, removed the data elements—date of birth and zip code—used by Sweeney[19] to identify the records of Massachusetts Governor William Weld from a publicly available database of state employee hospital records, believed to be the first published incident of a re-identification of a healthcare database presumed to be anonymous.[20]

Concerns about the de-identification standard were increased by highly publicized re-identifications of presumed 'anonymous' personal information posted on the internet.[21 22] The information in these incidents was not required to meet HIPAA de-identification standards nor any other legal or agreed-upon scientific standard requiring a very low risk of re-identification. These incidents, in combination with Sweeney's previous re-identification work, were cited by Ohm in a 2010 article concluding that 'in the past 15 years, computer scientists have established… the *easy re-identification result*', proving that the notion that data can be de-identified to zero privacy risk is 'deeply flawed' and should be rejected as a 'privacy-providing panacea'.[20]

More recent articles have challenged the premise of the 'easy re-identification result', at least with respect to data de-identified by HIPAA standards. For example, a 2011 review by El Emam *et al*[23] of 14 published re-identification attacks revealed that only six of the attacks were on health data, and only one of those was on data de-identified by HIPAA or similarly rigorous standards. The attack on HIPAA de-identified data had a very low re-identification rate of 0.013%.[23]

Other recent studies have focused specifically on the sufficiency of the safe harbor method, which presumes that removal of 18 specific data elements will continue to ensure a very small risk of re-identification, notwithstanding an ever-changing data ecosystem. One study focused on the sufficiency of the safe harbor method in light of the potential availability of voter registration data for linking purposes. (Sweeney used voter registration records to find Governor Weld in the hospital dataset.) The study concluded that the re-identification risk of the safe harbor method was likely to vary by location (due to differences in population distributions of US states), the potential linking variables contained within voter registries of various states, and varying access policies for voter registries.[24] The safe harbor has also been criticized as providing insufficient protection for datasets containing information at higher risk for re-identification but currently not required to be removed (such as some genetic information, longitudinal data, transactional data such as diagnosis codes, and free-form text).[25]

HHS in 2010 commissioned a study of the re-identification risk of a dataset compliant with the safe harbor. The study, involving admission records of Hispanic individuals in one hospital system between 2004 and 2009 and a hypothetical intruder with access to substantial market research information, re-identified only two of 15 000 individuals.[26] This re-identification was the sole successful example of re-identification in a HIPAA de-identified dataset identified in the 2011 review by El Emam *et al*.[23]

Concerns about HIPAA de-identification were raised by interested parties in the 2011 US Supreme Court case of Sorrell v.

Box 1 Current HIPAA methodologies for de-identification and limited datasets

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 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'.[14] The safe harbor option requires the removal of 18 specific data elements that could uniquely identify an individual, including names, all elements of dates more specific than a year, and most address information (except the initial three digits of a zip code in certain circumstances).[15] In addition, the data holder must not have 'actual knowledge' that the information in the de-identified dataset could be used to identify an individual subject.[16]

The safe harbor provides the easiest and most certain path to de-identification; however, because it requires the removal of precise dates and specific geographical information, it is often less useful for certain secondary purposes, such as health services research and syndromic surveillance. To respond to researchers concerns that the safe harbor standard resulted in data of limited utility for research purposes, the privacy rule was amended in 2002 to provide for the use of a limited dataset for healthcare operations, research or public health purposes.[17] To quality as a limited dataset, 16 categories of identifiers must be removed[18]; however, identifiers often deemed important for healthcare research, such as full dates and more specific geographical information, may be retained. Such information is considered to be 'identifiable' and therefore is regulated by the privacy rule; however, it can be accessed, used or disclosed for these purposes without the need to obtain subject consent or authorization (or a waiver thereof).[18] The recipient of the data is required to execute a data use agreement that sets the parameters for use of the data and prohibits re-identification of the subjects.[18]

IMS Health, Inc.[27] The case involved a challenge to a Vermont statute prohibiting the use of de-identified data for the marketing of brand name drugs. The statute's primary aim was controlling healthcare costs driven by the prescription of branded drugs. The state also claimed the need to protect patient privacy, although the statute relied solely on HIPAA de-identification to protect patient identities. In briefs submitted to the court, some academics and privacy advocates criticized the HIPAA de-identification standard and some of the de-identification techniques purportedly used by the data mining companies involved in the case; they urged the court to uphold the Vermont statute as an important patient confidentiality measure.[28] The court declined to use the case as an opportunity to critique HIPAA de-identification, instead declaring the Vermont statute to be unconstitutional because it targeted marketing uses of data by pharmaceutical manufacturers, violating their commercial free speech rights.[29] Some harshly criticized the decision for its failure to address the privacy of the doctor–patient relationship.[30 31]

THE VALUE OF DE-IDENTIFIED DATA

Notwithstanding concerns about the de-identification standard, it is critically important that HIPAA and other health privacy laws maintain a distinction between fully identifiable and de-identified data.[7 32] If privacy laws do not recognize this distinction, there will be no incentive for entities to expend resources to de-identify data and learn to work with them or to improve de-identification techniques. Instead, there will be a tendency to use fully identifiable data for secondary purposes when it is legally permissible, such as for public health and quality improvement, raising far greater privacy risk for individuals. In addition, pressure will increase to make identifiable data available to meet commercial data needs that currently rely on de-identified data. Re-identification may still be possible with de-identified data—but when de-identification is done properly, re-identification should not be easy or cheap.[24 33]

Now is the time to consider appropriate policies to address concerns about HIPAA de-identification. Delaying response until after a well-publicized re-identification could lead to policy that is more reactionary than reflective. HHS recognized in establishing the de-identification standard that there would still be some risk of re-identification.[13] Consequently, establishing policy that regulates even the low risk of re-identification makes sense.

POLICIES TO BUILD TRUST IN DE-IDENTIFIED DATA

The workshop convened by CDT in October 2011 included companies engaged in creating and/or using HIPAA de-identified data, academic experts on statistical de-identification, healthcare lawyers, and consumer advocates. CDT selected attendees because of their scholarship on de-identification, their experience in de-identifying data, their involvement in de-identification policy issues, and their interest in preserving privacy-protective mechanisms for using data to improve individual and population health. At the workshop, CDT gathered feedback on the following potential policy options, which were initially put forth by CDT in a 2009 white paper[7]:

1. Prohibiting unauthorized re-identification of de-identified data;
2. Ensuring the robustness of de-identification methodologies;
3. Establishing reasonable security safeguards for de-identified data;
4. Increasing public transparency about uses of de-identified data.

Each is discussed in more detail below.

Prohibiting unauthorized re-identification

Although there is currently little publicly available evidence that re-identification of HIPAA de-identified datasets is a common occurrence, the potential for re-identification—and the lack of accountability for those who do it—will be a persistent concern that, if not addressed, could create obstacles to more widespread uses of de-identified data. One solution is to hold individuals and entities legally accountable for unauthorized re-identification of de-identified datasets. Such policies would need to apply to recipients of de-identified data who are not HIPAA covered entities; they are permitted to re-identify and use health data consistent with the privacy rule.[34]

A few workshop participants expressed concern that prohibiting re-identification would force such activity further underground, making it more difficult to detect.[35] However, a number of workshop attendees favored policies establishing accountability for unauthorized re-identification in order to build public trust in uses of de-identified data. The Institute of Medicine has also recommended that re-identification without authorization be subject to legal sanctions.[36]

Accountability for unauthorized re-identification can be accomplished in the following two ways: (1) through legislation prohibiting recipients of de-identified data from unauthorized re-identification of the information; and (2) by requiring HIPAA-covered entities (and business associates) to obtain agreements with recipients of de-identified data that prohibit the information from being re-identified without authorization.

Both options are likely to require action by Congress, as HHS believes HIPAA does not give the department the power to regulate information that is not individually identifiable.[13] Furthermore, HIPAA coverage does not extend beyond covered entities and entities performing services on their behalf.

The first option has the advantage of potentially achieving more widespread coverage, including of health databases that may not currently be required to meet HIPAA de-identification standards and public use datasets, when acquiring enforceable agreement of data recipients not to re-identify may be a challenge. However, the second option may be easier to implement, as some workshop attendees noted that covered entities frequently already require de-identified data recipients to agree contractually not to re-identify.

Contractual provisions can be effective when the contracting parties choose to enforce them; however, it is also possible to create a 'hybrid option' in which contracts can be enforced by regulators or third parties. For example, Gellman[37] has developed model legislation, the Personal Data Deidentification Act, which would allow parties to a de-identified data agreement to opt into having it subject to enforcement by authorities or even by data subjects.

Legal prohibitions against re-identification may need to include exemptions to accommodate re-identification research—ie, attempts to re-identify intended to test how effectively a dataset is 'de-identified'—or possibly to allow re-identification of individuals for urgent health reasons (although covered entities de-identifying data already may include a re-identification code that they can use for this purpose). Such exemptions may be difficult to draft in legislation, and may need to be managed through the issuance of regulations or guidance by HHS. The contractual option would allow the parties to specify the narrow instances when re-identification would be permitted; however, to enable

consistent national policy on this issue, legislation or regulation could set more clear rules on when re-identification is permissible, and by whom.

Any law prohibiting re-identification will need to include a clear definition of that term; it may also be possible to protect re-identification research through a carefully crafted definition that focuses on actually identifying individuals. Gellman[37] defines re-identification as 'the linkage of deidentified personal information with an overt identifier which belongs or is assigned to a living or dead individual'. One workshop attendee suggested re-identification could be defined as an attempt to link data to categories of identifiers required to be excluded as part of a HIPAA limited dataset (see box 1 for an explanation of a limited dataset).

Ensuring robustness of de-identification methodologies

HHS created the methodologies for de-identification in regulation, so concerns about their sufficiency can be addressed without further legislation. Workshop participants largely agreed that strengthening these methodologies—coupled with accountability for re-identification—could considerably ease concerns about de-identification.

Since its inception the safe harbor methodology has been criticized for, among other things, failing to account for data recipients' potential access to information that can be used to re-identify and their motivation to re-identify. But HHS has rejected calls to rely solely on the statistical methodology, noting that they intended to create an easy to follow, 'cookbook' approach to de-identification, which could be used by entities without access to a statistician: '[t]he Safe Harbor is intended to involve a minimum of burden and convey a maximum of certainty that the rules have been met…'.[13] Of note, when the safe harbor method was initially proposed, covered entities were required to have 'no reason to know' the recipient could potentially re-identify the data in order for it to qualify for safe harbor status. However, covered entities did not want to be legally liable for failing to estimate correctly what linking information a data recipient might be able to access. Consequently, in the final rule, HHS took the guesswork out of the standard and required only that covered entities not have 'actual knowledge' of re-identification possibilities.[13]

It is unlikely HHS would agree to eliminate the safe harbor; furthermore, its ease of use and policy imperatives to encourage data to be used in least identifiable form counsel against elimination. The more palatable option is for HHS to review the safe harbor standard regularly, such as biennially, to ensure it continues to provide a very low risk of re-identification. In addition, if the current safe harbor proves to be vulnerable in certain contexts, its use could be precluded in those contexts.

The statistical methodology, which at least requires express consideration of other information that the recipient may be able to use to re-identify individuals, has also been criticized. Its viability depends on the quality of the statistical analysis, and there are currently no independent, objective mechanisms for vetting statistical analyses. Some have also argued that the 'very small risk' standard is too vague.[13]

In the final privacy rule, HHS recognized that entities choosing the statistical method of de-identification might need guidance to help them confidently achieve the 'very small risk' standard. HHS expressly listed two sources on statistical disclosure of confidential information published by the US Office of Management and Budget.[13] HHS acknowledged that for guidance on statistical techniques to be valuable, HHS would need to update it to keep up with technology and the availability of public information from other sources.[13] HHS committed to providing said such updated guidance in the future, but as of May 2012, it has not done so.

A number of workshop attendees supported further exploration of the following policy options, aimed at strengthening both de-identification methodologies:

- HHS could create a process for objectively vetting or setting standards for the statistical methodology, to provide some assurance to the public that the methodologies meet the very low risk standard. Most were supportive of having the techniques and practices used in the statistical methodology vetted by the federal government, using statistical experts at the National Institute of Standards and Technology, the National Center for Health Statistics, and/or the Census Bureau, to establish trust and a level playing field. However, private sector entities that operate with full transparency, objectivity, and accountability might aptly fill the vetting role.
- The current safe harbor standard should be reviewed on a regular basis. Such review could also be done by the 'objective vetters' suggested above to bolster the statistical methodology. In addition, safe harbor status—and its legal certainty as a de-identification methodology—could be extended to those statistical methodologies that satisfy the objective vetting process described above. This has the potential of adding reliable recipes to the de-identification 'cookbook', increasing their use and potentially reducing the cost of statistical de-identification. Although the current safe harbor applies to all recipients for all use scenarios, it is possible that safe harbor status should be granted only in circumstances in which a methodology has been demonstrated to achieve the very low risk standard. Any new methodology blessed with safe harbor status would also need to be reviewed on a consistent basis.
- HHS could also explore certifying or accrediting entities that regularly de-identify data or evaluate re-identification risk; those whose methodologies pass the objective vetting criteria established above would then be deemed as certified or accredited. Such 'centers of de-identification excellence' would need to be re-certified or accredited on a regular basis. Again, the 'objective vetters' deployed to review statistical and safe harbor de-identification methodologies could play a role in designing and potentially implementing or overseeing the implementation of a certification or accreditation system. Such certification/accreditation could begin as a voluntary initiative, on the theory that most health data mining companies would seek it to demonstrate trustworthiness to customers and the public; mandates could be imposed if voluntary initiatives fail or when circumstances require a higher level of trust.
- HHS should consider whether strengthening the safe harbor standard is sufficient to protect information in public use datasets. This could be particularly important if effective re-identification prohibitions for these data are not achievable. If re-identification risk depends on the motivation of the data recipient, and potential access to other information that can facilitate re-identification, it is more difficult to consider those risks with a dataset open to the public.

Workshop attendees provided feedback on a few other ideas. For example, should there be retention limits on de-identified datasets or a requirement to refresh them after a period of time to help ensure they continue to raise a very low risk of re-identification? Furthermore, some thought HHS should explore creating mathematical standards establishing what constitutes 'very small risk' of re-identification for different datasets and different purposes. However, others expressed concern that such

a standard would be impossible to calculate reliably, given that re-identification risk is contextual. Both of these ideas, as well as the recommendations above, require further exploration before being implemented as policy.

Reasonable security safeguards

Workshop attendees generally agreed with the idea that reasonable information security safeguards should protect all health information, and such safeguards should be commensurate with the privacy risks posed by the data. Consequently, in the case of de-identified data, the degree of security required need not be as robust as that needed to protect identifiable data or data at greater risk of re-identification. For example, the HIPAA security rule requires protections for all electronic protected (identifiable) health information.[38]

If security safeguards should be commensurate with the risk posed by the data, data holders probably need some flexibility to determine appropriate security measures to adopt. At a minimum, holders of de-identified data should be held accountable for adopting security measures that protect against prohibited re-identification or ensure that commitments can be honored with respect to re-identification or limiting the particular uses of de-identified data. For example, if de-identified data are released by a covered entity for research purposes only, the recipient should adopt appropriate physical and technical safeguards to ensure access is limited to those conducting the research. Implementing security safeguards for public use datasets will be a particular challenge.

Transparency to the public

Distinct from concerns about the potential risks of de-identified data to confidentiality are concerns about actual uses of de-identified data. The Vermont statute in Sorrell restricted the use of de-identified data (see box 2) for pharmaceutical marketing purposes; similar statutes had been enacted in Maine and New Hampshire.[27] Concerns have also been raised about de-identified data informing business decisions in ways that could have a negative impact on patients.[39] FICO recently launched a 'medication adherence score' tool that purports to use de-identified data to predict whether patients will adhere to medication regimens.[40] Officials from FICO claimed the information could not be used by insurers for risk adjustment purposes[41]; nevertheless, the news alarmed a number of consumer advocates.[42 43]

Many uses of de-identified data currently occur with little public knowledge, and this lack of transparency contributes to public concerns about health data de-identification.[7 35] Greater transparency about the de-identification standard, as well as on uses (and users) of de-identified data, could help build trust in uses of de-identified data; it could also help uncover concerns about uses of de-identified data that may be addressable through more direct policy. However, workshop attendees noted that effective outreach to the public on this issue will be a challenge; individuals often do not have good knowledge even of the uses and disclosures of identifiable health data.[50] HHS could consider funding pilot projects to increase transparency of de-identified data or tying federal funding or favorable regulatory treatment—such as safe harbor status for de-identification methodologies or Center of De-Identification Excellence status—to greater public transparency about de-identification and uses of de-identified data.

CONCLUSION

Increasing concerns about re-identification risks could erode trust in the HIPAA de-identification standard. But de-identification, if done correctly, provides an important tool for privacy protection while preserving data utility for uses critical to advancing a more effective and efficient healthcare system. Policies to reduce the risk of re-identification, encourage use of strong de-identification methods and practices, and enhance public awareness of uses of de-identified data could help ease concerns and build public trust in a more robust health data ecosystem.

Box 2 Should individuals have the right to consent to uses of de-identified data?

Even in circumstances in which the information raises a very low risk of re-identification, some individuals have heightened sensitivity regarding uses of health information about them.[44] A 2010 survey by the California Healthcare Foundation found that 'a majority of adults express discomfort (42 per cent) or uncertainty (25 per cent) with their health information being shared with other organizations — even if… [their] name, address, [date of birth and social security number] were not included'.[45] Some have suggested that individuals should have the right to consent to—or at least be able to opt out of—having their data included in de-identified datasets.[35 39]

The surveys are not consistent on this issue. A survey released by the Markle Foundation in 2011 found that at least 68% of the public, and 75% of doctors expressed willingness to 'allow composite information to be used to detect outbreaks, bio-terror attacks, and fraud, and to conduct research and quality and service improvement programs'.[46] Markle noted that this result was consistent with a similar survey it conducted in 2006.

Many workshop attendees expressed concern that allowing individuals to consent (either opt in or out) to being included in de-identified datasets treats de-identified data as though they raise the same risk as identifiable data, providing a disincentive to de-identify. In addition, consent in practice too often provides weak privacy protection, suggesting that relying on it as a mechanism to give individuals a voice in how de-identified data are used would probably not be very effective.[47 48] Even more importantly, before implementing any policies requiring consent for uses of de-identified data, policy makers should consider the literature exploring whether consent requirements introduce selection biases into de-identified datasets, potentially distorting the accuracy of results, increasing the costs, and lengthening the time for conducting scientific research and healthcare quality assessments with de-identified data.[49]

Promoting greater transparency of uses of de-identified data may be a more promising way to build public trust in de-identified data uses.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Safran C,** Bloomrosen M, Hammond WE, *et al*; Expert Panel. Toward a national framework for the secondary use of health data: an American medical Informatics Association white paper. *J Am Med Inform Assoc* 2007;**14**:1—9.
2. **PricewaterhouseCoopers LLP.** *Transforming Healthcare Through Secondary Use of Health Data*. 2009. http://www.pwc.com/us/en/healthcare/publications/secondary-health-data.jhtml (accessed 4 Mar 2012).
3. *Code of Federal Regulations* Title 45 *Subpart E (Sections 164.500-534)*. Washington, DC: Office of the Federal Register.

4. *Code of Federal Regulations* Title 45 *Section 164.502(d)(2)*. Washington, DC: Office of the Federal Register.
5. *U.S. Code* Title 42 *Section 17954(c)*. Washington, DC: Office of Law Revision Counsel of the U.S. House of Representatives.
6. **US Department of Health and Human Services.** *Workshop on the HIPAA Privacy Rule's De-Identification Standard*. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/deidentificationworkshop2010.html (accessed 5 Mar 2012).
7. **Center for Democracy & Technology.** *Encouraging the Use of, and Rethinking Protections for, De-Identified (and "Anonymized") Health Data*. 2009. https://www.cdt.org/paper/encouraging-use-and-rethinking-protections-de-identified-and-anonymized-health-data (accessed 4 Mar 2012).
8. **The White House.** *Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy*. 2012. http://www.whitehouse.gov/blog/2012/02/23/we-can-t-wait-obama-administration-calls-consumer-privacy-bill-rights-digital-age (accessed 4 Mar 2012).
9. **US Department of Health and Human Services.** *Health Insurance Portability and Accountability Act of 1996. Public Law 104—91*. Sections 1171-1172. http://www.hhs.gov/ocr/privacy/hipaa/administrative/statute/index.html (accessed 4 Mar 2012).
10. *Code of Federal Regulations* Title 45 *Sections 160.103 and 164.502(e)*. Washington, DC: Office of the Federal Register.
11. *U.S. Code* Title 42 *Section 1171(6)*. Washington, DC: Office of Law Revision Counsel of the U.S. House of Representatives.
12. *Code of Federal Regulations* Title 45 *Section 164.514(a) (Emphasis Added)*. Washington, DC: Office of the Federal Register.
13. **US, Department of Health and Human Services.** *Standards for Privacy of Individually Identifiable Health Information; Final Rule*. 65 Federal Register 82462-82829. 2000. 45 CFR Parts 160 and 164.
14. *Code of Federal Regulations* Title 45 *Section 164.514(b)*. Washington, DC: Office of the Federal Register.
15. *Code of Federal Regulations* Title 45 *Section 164.514(b)(2)(i)*. Washington, DC: Office of the Federal Register.
16. *Code of Federal Regulations* Title 45 *Section 164.514(b)(2)(ii)*. Washington, DC: Office of the Federal Register.
17. **US Department of Health and Human Services.** *Standards For the Privacy of Individually Identifiable Health Information; Final Rule*. 67 Federal Register 53182-53273. 2002. 45 CFR Parts 160 and 164.
18. *Code of Federal Regulations* Title 45 *Sections 164.512(e)(1)-(5)*. Washington, DC: Office of the Federal Register.
19. **Sweeney L.** Guaranteeing anonymity when sharing medical data, the Datafly System. *Proc AMIA Annu Fall Symp* 1997:51—5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233452/?tool=pubmed (accessed 5 Mar 2012).
20. **Ohm P.** Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA Law Rev* 2010;**57**:1701—77.
21. **Barbara M,** Zeller T Jr. A face is exposed for AOL searcher no. 4417749. *New York Times* August 2006.
22. **Narayanan A,** Shmatikov V. Robust de-anonymization of large datasets. Proceedings of the 29th IEEE Symposium on Security and Privacy, Oakland, CA, IEEE Computer Society, May 2008:111—25.
23. **El Emam K,** Jonker E, Arbuckle L, *et al*. A systematic review of re-identification attacks on health data. 2011;**6**. http://dx.plos.org/10.1371/journal.pone.0028071 (accessed 4 Mar 2012).
24. **Benitez K,** Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc* 2010;**17**:169—77.
25. **El Emam K.** Methods for the de-identification of electronic health records for genomic research. *Genome Med* 2011;**3**:25.
26. **Kwok P,** Lafky D. *Harder Than You Think: A Case Study of Re-identification Risk of HIPAA-Compliant Records*. 2011. http://www.amstat.org/meetings/jsm/2011/onlineprogram/AbstractDetails.cfm?abstractid=302255. (abstract only; copy on file with author). http://www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf (accessed 4 Mar 2012).
27. *Sorrell v. IMS Health Inc., 131 S. Ct. 2653*. 2011. http://www.scotusblog.com/case-files/cases/sorrell-v-ims-health-inc/ (accessed 4 Mar 2012).
28. *See Briefs of Amicus Curiae Submitted by the Electronic Frontier Foundation and by the Electronic Privacy Information Center (EPIC) and Legal Scholars and Technical Experts*. http://www.scotusblog.com/case-files/cases/sorrell-v-ims-health-inc?wpmp_switcher-desktop (accessed 4 Mar 2012).
29. **McGraw D.** *Lack of Genuine Privacy Interest Doomed Vermont Drug Marketing Law*. 2011. http://www.ihealthbeat.org/perspectives/2011/lack-of-genuine-privacy-interest-doomed-vermont-drug-marketing-law.aspx (accessed 4 Mar 2012).
30. **Singer N.** Data privacy, put to the test. *New York Times* 30 April 2011.
31. **Congressman Ed Markey.** http://markey.house.gov/press-release/july-8-2011-resolution-disapproval-supreme-court-decision-sorrell-v-ims-health-case (accessed 4 Mar 2012).
32. **Cavoukian A,** El Emam K. *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy*. 2011. http://www.ipc.on.ca/images/Resources/anonymization.pdf (accessed 4 Mar 2012).
33. **Yakowitz J,** Barth-Jones D. *The Illusory Privacy Problem in Sorrell v. IMS Health*. 2011. Technology Policy Institute. http://www.techpolicyinstitute.org/news/show/23297.html (accessed 4 Mar 2012).
34. *Code of Federal Regulations* Title 45 *Sections 164.502(d)(2)(i) & (ii)*. Washington, DC: Office of the Federal Register.
35. **Data Privacy Lab.** *Comments of Latanya Sweeney and the Data Privacy Lab, Submitted to the Department of Health and Human Services Re: The Advance Notice of Proposed Rulemaking: Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators*. Docket ID number HHS-OPHS-2011-0005. 2011. http://dataprivacylab.org/projects/irb/DataPrivacyLab.pdf (accessed 4 Mar 2012).
36. **Institute of Medicine.** *Beyond the HIPAA Privacy Rule, Enhancing Privacy, Improving Health Through Research*. Washington, DC: National Academies Press, 2009:8.
37. **Gellman R.** The deidentification dilemma: a Legislative and contractual proposal. Fordham Intellectual Property. *Media & Entertainment Law Journal* 2010;**21**:33—61.
38. *Code of Federal Regulations* Title 45 *Part 164 Subpart C (Sections 164.302-318)*. Washington, DC: Office of the Federal Register.
39. **Rothstein MA.** Is de-identification sufficient to protect health privacy in research? *Am J Bioeth* 2010;**10**:3—11.
40. **Business Wire.** *New FICO Analytis Predict Likelihood of Patient Adherence to Prescription Medication*. 2011. http://www.businesswire.com/news/home/20110623005735/en/FICO-Analytics-Predict-Likelihood-Patient-Adherence-Prescription (accessed 4 Mar 2012).
41. **Creditcards.com.** http://www.creditcards.com/credit-card-news/fico-score-medication-adherence-1270.php (accessed 4 Mar 2012).
42. **ctwatchdog.com.** http://ctwatchdog.com/health/medical-privacy-issue-fico-medication-adherence-score-coming (accessed 4 Mar 2012).
43. **Daily Kos.** http://www.dailykos.com/story/2011/06/21/987286/-Scary-stuff:-FICO-scoring-millions-of-Americans-on-medication-compliance (accessed 4 Mar 2012).
44. **Allen A.** *Privacy and Medicine. 2009 Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/privacy-medicine (accessed 4 Mar 2012).
45. **California Healthcare Foundation.** *Consumers and Health Information Technology: A National Survey*. 2010. http://www.chcf.org/publications/2010/04/consumers-and-health-information-technology-a-national-survey (accessed 4 Mar 2012).
46. **Markle.** *The Public and Doctors Overwhelmingly Agree on Health IT Priorities to Improve Patient care*. 2011. http://www.markle.org/publications/1461-public-and-doctors-overwhelmingly-agree-health-it-priorities-improve-patient-care (accessed 4 Mar 2012).
47. **Cate F.** Protecting privacy in health research: the limits of individual choice. *Calif Law Rev* 2010;**98**:1765—804.
48. **McGraw D,** Dempsey JX, Harris L, *et al*. Privacy as an enabler, not an impediment: building trust into health information exchange. *Health Aff (Millwood)* 2009;**28**:423—4.
49. **El Emam K,** Dankar FK, Issa R, *et al*. A globally optimal k-anonymity method for de-identification of health data. *J Am Med Inform Assoc* 2009;**16**:670—82.
50. **Westin A.** *National Partnership for Women & Families. Making IT Meaningful: How Consumers Value and Trust Health IT*. 2012. http://www.nationalpartnership.org/site/PageServer?pagename=issues_health_IT_survey (accessed 5 Mar 2012).