KEEPING THE INTERNET
OPEN • INNOVATIVE • FREE

CENTER FOR DEMOCRACY & TECHNOLOGY

De-Identification of Health Data under HIPAA: Potential Paths Forward

Deven McGraw
Director, Health Privacy Project
November 30, 2011

Health Privacy Project at CDT

- Project's aim: Develop and promote workable privacy and security policy solutions for personal health information.
- Workable means privacy as *enabler*. Privacy is not the endpoint – it is the means to building trust in data sharing, including for secondary purposes.
- Without privacy protections, people will engage in "privacy-protective behaviors" to avoid having their information used inappropriately.
- Public trust in electronic data infrastructures depends on keeping data confidential & protected from unauthorized access, use and disclosure

Federal (HIPAA) Policy on "De-identification"

- "De-identified data" = data that meets HIPAA standard for deidentification
- Data that meets the HIPAA de-identification standard is not PHI and largely not regulated by HIPAA
- De-identification standard = no reasonable basis to believe the data can be used to identify an individual (45 CFR 164.514(a))

De-identified Data under HIPAA

- Two methods may be used to de-identify:
  - Statistical method requires someone with statistical expertise to determine that the risk is **very small** that the information, on its own or in combination with other reasonably available information, could be used by an anticipated recipient to identify an individual (164.514(b)(1))
  - Safe harbor requires the removal of 18 specific data elements; in addition, data holder must not have **actual knowledge** that the data, either alone or in combination with other data, could identify an individual.(164.514(b)(2))
- Entities may assign a code to allow de-identified data to be re-identified as long as code is not shared. (164.514(c))

De-identification Policy Challenges (1)

- Risk of re-identification is contextual:
  - What other data does the data recipient have access to
  - What is the recipient's motivation to re-identify or use inappropriately
- HIPAA safe harbor approach assumes a static environment and automatically concludes that data can be deemed to raise a very small risk without consideration of this context

De-identification Policy Challenges (2)

- Ensuring very small risk of re-identification – particularly through safe harbor standard – could get more difficult over time, due to increased availability of data
- Statistical method for de-identification is meant to be flexible over time – but robustness depends on quality of statistical analysis.
- De-identification means very small risk, not no risk - we still need policies to address the risk that does exist
- Concerns about uses of de-identified data, and the robustness of de-identification methodologies, appears to be increasing (*Sorrell v. IMS Health Inc. et al*.)

Less Identifiable = Less Risk

- Even if there are limits to whether true de-identification can be achieved, this does not mean all data present equal risk
- De-identifying or removing identifiers from data, or shielding identity through use of technology, provides additional protections for confidentiality and preserves utility
- HIPAA requirements to use the minimum necessary amount of information needed to accomplish a particular purpose arguably applies to data identifiability

cdt KEEPING THE INTERNET OPEN • INNOVATIVE • FREE

Is it Time to Strengthen our De-identification Policies?

- Yes (CDT's view)
  - Concerns about de-identification based on misinformation could drive bad policy.
  - Gaps exist that could be fixed without the need for significant overhaul.
  - Need approach that ideally has some consistent threads with proposed policy approaches to anonymization on the Internet
- Contrary view
  - No evidence that system is broken – so no need to fix.
  - Opening this issue could still result in bad policy (too often politics, not substance, drive policymaking).

cdt KEEPING THE INTERNET OPEN • INNOVATIVE • FREE

CDT's previous work on this issue

- Initial workshop on de-identification held in 2008; white paper of potential policy solutions released in 2009.) "Encouraging the Use of, and Rethinking Protections for, De-Identified (and "Anonymized") Health Data": http://www.cdt.org/files/pdfs/20090625_deidentify.pdf
- Held follow-up workshop in October 2011 to drill down with more specificity on some of the policy solutions initially floated in the 2009 paper.

Potential paths forward - intro

- Dealing with potential for re-identification, through statutory prohibition or requiring contractual commitments.
- Assuring robustness of HIPAA de-identification methodologies
  - Review of (and addition to) safe harbor
  - Certification or recognition of statistical methods
- Reasonable security safeguards
- Greater public transparency about de-identified data uses.

Prohibiting re-identification (via statute or contract)

- Con
  - No evidence that re-identification is occurring – trying to craft a workable prohibition could do more harm
  - Would need exceptions to continue to allow individuals to "test" robustness of de-identified data sets and to be able to notify individuals of important information about their health
  - Fears of liability (real or perceived) could make de-identification more expensive (through overly conservative behavior, for example)
- Pro
  - Taking prospect of re-identification "off the table" could increase public confidence in de-identification, uses of de-identified data
  - Reduces motivation to re-identify, which reduces re-identification risk

cdt KEEPING THE INTERNET OPEN • INNOVATIVE • FREE

Assuring Robustness of HIPAA De-identification methodologies

- In general, high degree of support for this at recent workshop – trick is figuring out the details
- Promising ideas we are pursuing:
  - Require HHS (with assistance from NIST) to establish standard(s) for acceptable risk of re-identification that ALL de-identification techniques must meet. Should be subject to review every X years.
  - HHS must evaluate current safe harbor methodology against this percentage every X years.
  - HHS should add methodologies to the safe harbor to provide other viable and effective options for de-identification – The safe harbor method is straightforward, and provides legal certainty for providers; more options that provide this level of assurance should be available.

cdt KEEPING THE INTERNET OPEN • INNOVATIVE • FREE

Assuring Robust Methodologies – More Promising ideas

- Limit use of current safe harbor (to better manage re-identification risk)
  - Cannot be used if data is not a random sample of US population
  - Cannot be used when certain fields that are highly susceptive to re-identification are present (genetic data, longitudinal data, ICD diagnosis codes, notes)
  - Cannot be sole de-identification methodology if information is made publicly available
  - Concern: how to deal with legacy safe harbor datasets

cdt KEEPING THE INTERNET OPEN • INNOVATIVE • FREE

Assuring Robust Methodologies – More Promising ideas (2)

- Create objective vetting process for entities using statistical methodologies (or combination of safe harbor and statistical methods)
  - Could be certification process, or recognition of "Centers of Excellence"
- Key Questions to resolve:
  - Who would vet? Government is one possible, neutral choice – but may not have sufficient expertise (could they develop it?)
  - Could a private sector process be developed over time? (NCQA, JCAHO)
  - What is the incentive/reward for achieving certification/Center of Excellence? (one possibility: add to safe harbor)

cdt KEEPING THE INTERNET OPEN • INNOVATIVE • FREE

Reasonable security safeguards

- Workshop participants seemed to agree, as long as safeguards commensurate with risk raised by the data
- What "reasonable" safeguards for de-identified data look like?
  - HHS Research office recently proposed that the minimum be a commitment not to re-identify (Advanced Notice of Proposed Rulemaking on the Common Rule)
  - Require more for safe harbor-only de-identified data sets? (but no evidence at this point they are more risky)
  - How to police for public use datasets?

cdt KEEPING THE INTERNET OPEN • INNOVATIVE • FREE

Greater transparency to the public

- All seemed to agree that greater public transparency – particularly re: beneficial uses of de-identified data – would be helpful for public policy. Insufficient time at workshop to generate specific ideas (just ask an advertiser – it's hard to reach the public effectively).
- Education would ideally be of all uses of de-identified data, so it's not the mystery "black box" (CDT theory: greater public transparency can help lead to greater public acceptance)
- Even greater challenge – dealing with data uses that may be objectionable to some
  - Possibility of bias if allow people to opt-in or out; regulating certain uses could be problematic (Sorrell).
  - Could IRBs play a role?

cdt KEEPING THE INTERNET OPEN • INNOVATIVE • FREE

CDT on "Deidentification"

- White Paper (June 2009): "Encouraging the Use of, and Rethinking Protections for, De-Identified (and "Anonymized") Health Data": http://www.cdt.org/files/pdfs/20090625_deidentify.pdf
- Policy Post (shorter version of above) (6/26/09): http://www.cdt.org/policy/stronger-protections-and-encouraging-use-de-identified-and-anonymized-health-data
- iHealthBeat Perspectives (even shorter) (7/30/09): http://www.ihealthbeat.org/perspectives/2009/anonymized-medical-data-protects-privacy-improves-care.aspx

Questions?

Deven McGraw

202-637-9800 x115

deven@cdt.org

www.cdt.org/healthprivacy