Is Sharing De-identified Data Legal? The State of Public Health Confidentiality Laws and Their Interplay with Statistical Disclosure Limitation Techniques

Victor Richardson, Sallie Milam, and Denise Chrysler

The diversity of state confidentiality laws governing public health data presents a significant challenge for public health initiatives. This challenge is further complicated by the array of confidentially laws that are relevant within a state as disclosure and usage standards vary depending upon data holder, type, and source. These laws often have not been updated to address modern confidentiality risks such as unlawful data linkage or breach, leaving many public health organizations without clear guidance in the contentious area of individual privacy. To address these challenges, public health organizations have increasingly turned to the science of de-identification, but whether de-identification adequately meets the many and varied state confidentiality legal requirements remains an unanswered question.

Understanding De-identification Science

Despite their diverse nature, most public health confidentiality laws share a common theme: the disclosure of personal health information is generally prohibited, but the disclosure of non-personal health information is not. Thus, if personal health information can be rendered non-personal, it may be freely disclosed.[1] The line between personal and non-personal, however, is not easily defined.[2] Even national obesity statistics are arguably personal; they describe the likelihood any given individual represented by the sample group was obese during the reporting period. Accordingly, statisticians refer to the science of rendering personal data impersonal as statistical disclosure *limitation.*[3]

Statistical disclosure limitation attempts to minimize the risk that a disclosure will reveal either the identity of, or information about, an individual. Initially, this science addressed only aggregate or statistical reports, but, as technology advanced, demand for the raw data behind the reports increased. Releasing this raw data allowed for greater flexibility and discovery, but also created a greater risk to confidentiality. Raw data is often comprised of a multitude of individual records and may include information that directly or indirectly identifies individuals. To manage this new risk, more sophisticated statistical disclosure limitation techniques were needed.[4]

In 2003, the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule took effect, establishing a national standard for protecting the confidentiality of health care data and addressing the challenges presented by the disclosure of record level data. Under HIPAA, data containing protected health information may be disclosed after it has been "de-identified" using one of two methods: safe harbor (where 17 specified identifiers are removed) and statistical de-identification (where a statistician determines that the "risk is very small that the information could be used…to identify an individual"). HIPAA also provides for the release of a "limited data set," which is not considered de-identified, for research purposes.[5]

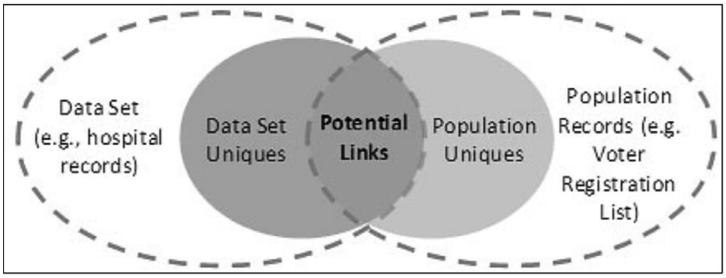
When the HIPAA privacy rule was being drafted, however, public health organizations, in recognition of their mission and long history of positive data stewardship, were exempted from many of its requirements, including its de-identification standards.[6] States were left free to implement and enforce their own

Victor Richardson, J.D., CIPP/US, *is a Robert Wood Johnson Foundation sponsored Public Health Law Fellow stationed at the West Virginia Healthcare Authority.* **Sallie Milam, J.D., CIPP/US/G,** *is the Chief Privacy Officer for the state of West Virginia at the WV Health Care Authority.* **Denise Chrysler, J.D.,** *is the Director for the Mid-States Regional Center of the Network for Public Health Law, located at the University of Michigan School of Public Health.*

de-identification regimes for public health data, but that flexibility had a price. Without prescribed standards, public health organizations are often unsure if and how they can share de-identified data under their state confidentiality laws. This uncertainty has been exacerbated by recent academic studies attempting to demonstrate it is possible to re-identify individuals in purportedly confidential data releases.[7]

The Re-identification Threat

Even when direct identifiers are removed from data, it is statistically possible for an attacker to re-establish identities by discovering pockets of uniqueness that remain. Certain combinations of values may be so unique that they serve as a "fingerprint" pointing to one individual. A re-identification attack attempts to locate the unique fingerprints in a purportedly confidential dataset and match those fingerprints to another dataset containing direct identifiers.[8] De-identification attempts to protect against this risk.



[9]

In recent years, researchers have studied techniques to re-identify purportedly confidential datasets. These studies often report startling high success rates, and have caused some scholars to question the efficacy of de-identification entirely.[10] For example, an often cited 2000 study found 87% of the U.S. population could be uniquely identified by their combination of gender, date of birth, and zip code.[11] Even when new researchers replicated the study to reflect a growing population, they still found 63% of the population uniquely identifiable using these variables.[12] Out of context, these numbers are startling.

In reality, however, these unique and exact combinations of gender, date of birth, and zip code would never be present in a de-identified dataset. Such combinations are either generalized or removed entirely, drastically reducing the risk of re-identification.[13] In the latter study above, researchers found the risk of unique identification dropped sharply when given slightly more abstract data. When they replaced an individual's full date of birth with only the month and year, only 4.2% of the population remained uniquely identifiable, and when they also replaced zip code with county, just 0.2% remained uniquely identifiable.[14] More impressively, data de-identified using the HIPAA safe harbor method is said to present only a .04% risk of unique identification.[15]

Still, the majority of re-identification studies continue to target data that is not truly de-identified, leading to what some call "the myth of easy re-identification."[16] While academics and scientists debate de-identification's merits, however, a more pertinent question has been neglected: is sharing de-identified data legal?

The Legality of De-identification Science

Few courts have considered the de-identification and disclosure of public health information in light of state confidentiality laws, but their decisions provide a window into the legality of sharing de-identified public health data across the country. The governing confidentiality standard differed in each case, but those differences proved largely irrelevant. The courts in question ultimately based their decision on a fact-based determination of whether the contested disclosure would place the confidentiality of personal health information at undue risk.

In *Southern Illinoisan v. Illinois Department of Public Health*, the Supreme Court of Illinois examined a newspaper's request for cancer registry information concerning neuroblastoma incidence by type of cancer, zip code, and diagnosis date. The Illinois Department of Public Health refused to provide this information, citing state confidentiality law that "precludes disclosure of…'[t]he identity, or any group of facts which tends to lead to the identity, of any person…'" and expert testimony that patients could be identified by matching the data requested with publicly available data.

The court did not find this argument persuasive, concluding that "information 'tends to lead to the identity' of Registry patients only if that information can be used by the general public to make those identifications," and noting that the expert methodology presented was "unique to [the expert's] education, training and experience…." Because the department did not produce any evidence that a member of the general public could perform the multi-step procedure to match identities, the court ordered the data be disclosed.[17]

In *Marine Shale Processors, Inc. v. State of Louisiana*, the First Circuit Court of Appeal of Louisiana considered a discovery request for data from a study, conducted by the Louisiana Department of Health and Hospitals, involving five cases of neuroblastoma found in St. Mary Parish. Narrower in scope than the data request in *Southern Illinoisan*, this data request also sought far more granular data, including diagnosis date, age at diagnosis, sex, race, religion, family medical histories, diagnostic information, treatment, and vital status of the patients. This granularity, in combination with the small population size, ultimately swayed the court to rule against disclosure. Of particular concern to the court was that, due to St. Mary Parish's unusually high incidence of neuroblastoma, the identities of the five subjects were already generally well known.[18]

> Although there are relatively few court decisions on point outside of scenarios involving small and well-known populations, courts have made it clear that they will not forbid disclosures absent real evidence that information can be readily used to re-identify individuals. Therefore, so long as de-identification science meets that burden, it will remain an effective and legal means to maintain the confidentiality of personal health information.

Later in 2004, the same Louisiana court provided additional guidance in *Williams Law Firm v. Board of Supervisors*. Here, a law firm requested annualized information concerning 21 rare cancers in seven parishes from the Louisiana Tumor Registry. The Registry refused to provide the information, relying on state law precluding disclosure of "case specific data, as distinguished from group, tabular, or aggregate data concerning patients…" to argue that it could not release incident rates of one or zero. The firm argued that such values would only be case-specific if the population of a given parish was one individual. The court found the firm's argument persuasive, holding that, in this case, values of zero or one were not identifiable or case specific.[19]

Thus, although there are relatively few court decisions on point outside of scenarios involving small and well-known populations, courts have made it clear that they will not forbid disclosures absent real evidence that information can be readily used to re-identify individuals. Therefore, so long as de-identification science meets that burden, it will remain an effective and legal means to maintain the confidentiality of personal health information.

Conclusion

The science of de-identification continues to advance, and data de-identification has become an accepted form of protecting the confidentiality of personal information under federal regulation. At the same time, re-identification studies have continued to focus on data disclosures that fail to meet any modern standard of de-identification. Thus, while public health organizations may lack specific guidance on how to de-identify data in a way permissible under their applicable state confidentiality laws, they can reasonably rely on the efficacy of modern de-identification techniques, so long as the governing confidentiality standard allows for the disclosure of data that does not identify an individual. At the same time, health organizations which intend to resist requests for disclosure should be prepared to demonstrate why certain data requests could well lead to the identification of specific personal information.

References

1. J. M. Ware, "Public Health Departments and State Patient Confidentiality Laws Map, Law Atlas: The Policy Surveillance Portal," *available at* <http://lawatlas.org/query?dataset=public-health-departments-and-state-patient-confidentiality-laws#.VEqz-vnF98F> (last visited February 5, 2015).
2. See, e.g., P. Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *University California Los Angeles Law Review* 57, no. 6 (2010): 1701-1777.
3. See, e.g., American Statistical Association: Committee on Privacy and Confidentiality, *Key Terms/Definitions* (2011), *available at* <http://community.amstat.org/CPC/AboutUs/KeyTermsDefinitions> (last visited February 5, 2015).
4. V. Ciriani et al., "Microdata Protection," in Ting Yu and Sushil Jajodia,, eds., *Secure Data Management in Decentralized Systems* (Springer, 2007): at 291-321, *available at* <http://spdp.di.unimi.it/papers/microdata.pdf> (last visited February 5, 2015).
5. *Guidance Regarding Methods for De-identificaiton of Protected Health Information in Accordance with the Health Insurance Portability and Accountability (HIPAA) Privacy Rule*, Department of Health and Human Services: Understanding HIPAA, *available at* <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html> (last visited February 5, 2015) [hereinafter cited as De-identification in Accordance with HIPAA].
6. Center for Disease Control and Prevention, "HIPAA Privacy Rule and Public Health: Guidance from CDC and the U.S. Department of Health and Human Services," April 11, 2003, *at* <http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm> (last visited February 5, 2015).
7. A. Cavoukian and K. E. Emam, *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy* (June 16, 2011), Discussion Papers, Information and Privacy Commissioner of Ontario, *available*

at <http://www.ipc.on.ca/images/Resources/anonymization.pdf> (last visited February 5, 2015).

8. De-identification in Accordance with HIPAA, *supra* note 5; Ciriani et al., *supra* note 4.
9. *Id.* (De-identification in Accordance with HIPAA).
10. See Ohm, *supra* note 2.
11. L. Sweeney, *Simple Demographics Often Identify People Uniquely*, Data Privacy, Carnegie Mellon University, Working Paper 3, Pittsburgh (2000), *available at* <http://dataprivacylab.org/projects/identifiability/paper1.pdf> (last visited February 5, 2015).
12. P. Golle, *Revisiting the Uniqueness of Simple Demographics in the US Population*, paper presentation at the 5th ACM Workshop on Privacy in Electronic Society, ACM, New York, New York, United States, 2006, *available at* <http://crypto.stanford.edu/~pgolle/papers/census.pdf> (last visited February 5, 2015).
13. See Cavoukian, *supra* note 7.
14. See Sweeney, *supra* note 11.
15. National Committee on Vital and Health Statistics, *Enhanced Protections for Uses of Health Data: A Stewardship Framework for 'Secondary Uses' of Electronically Collected and Transmitted Health Data*, Report to the Secretary of the U.S. Department of Health and Human Services (December, 19, 2007), *available at* <http://www.ncvhs.hhs.gov/071221lt.pdf> (last visited February 5, 2015).
16. See Cavoukian and Emam, *supra* note 7.
17. *Southern Illinoisan v. Illinois Dep't of Public Health*, 844 N.E.2d 1 (Ill. 2006).
18. *Marine Shale Processors, Inc. v. State of Louisiana Dep't of Health*, 572 So. 2d 280 (La. App. 1 Cir. 1990).
19. *Williams Law Firm v. Board of Supervisors*, 878 So. 2d 557 (La. App. 1 Cir. 2004).