The ability to generate and use large amounts of information about patient and population health, health-care treatment, and the outcomes of care represents a fundamental breakthrough in health-care quality improvement and population health management. Creating and using what has become known as "big data" raises many legal issues, as this installment of *Law and the Public's Health* describes. This article was published on January 26, 2015, at www.publichealthreports.org.

Sara Rosenbaum, JD
George Washington University, Milken Institute School of Public Health
Department of Health Policy, Washington, DC

BIG DATA AND PUBLIC HEALTH: NAVIGATING PRIVACY LAWS TO MAXIMIZE POTENTIAL

Jane Hyatt Thorpe, JD
Elizabeth Alexandra Gray, JD

This installment of *Law and the Public's Health* examines the constellation of laws governing health information privacy and their application to big data in public health. Big data holds great promise for public health, given the nature of services such as monitoring population health status, evaluating population-based health service quality, and conducting research for innovative solutions.[1] These activities require the ability to collect volumes of information, rapidly interpret data, and monitor data for long periods of time—all functions that are the hallmark of big data solutions. Despite misconceptions about health information privacy laws, the legal framework is quite permissive and need not operate as a barrier.

This article defines big data, provides an overview of how laws related to health information privacy apply to big data, and discusses the implications of this framework for public health policy and practice.

BACKGROUND

"Big data" is characterized by three Vs—volume, velocity, and variety—and cannot be managed with standard database processing methods.[2] Big data's value is in its use—combining large amounts of information from multiple sources into a single dataset permits identification of correlations and patterns that would be hidden in siloed datasets.[3] Access to a single big dataset allows users to fill in information gaps and check for consistency, capturing a more accurate picture of the population being evaluated.[4] Big data technologies also permit the storage of volumes of information indefinitely so that data can be used in the future for a purpose other than that for which they were originally collected (i.e., a secondary use).[5] The University of Pittsburgh is demonstrating the value of secondary uses through Project Tycho, which put 88 million disease reports published since 1888 in the Centers for Disease Control and Prevention's (CDC's) *Morbidity and Mortality Weekly Report* into an open-access database.[6] The team used the database to demonstrate the effect of vaccinations on diseases such as polio, and other uses neither considered nor possible when the information was first collected.

Solutions for data collection, storage, processing, and analysis have become abundant and affordable, permitting users across industries to manage the size, speed, and complexity of big data.[1] These technological advancements have ushered in what some herald as the "big data revolution,"[7] although its full potential in the health-care sector has not been realized due to financing, interoperability issues, and legal concerns related to information privacy and security.

HEALTH INFORMATION PRIVACY AND THE LEGAL FRAMEWORK FOR BIG DATA

The health information privacy framework is a patchwork of often-overlapping federal and state laws that regulate specific types of information, individuals, and organizations.[8] Confusion related to the scope and application of this framework, as well as the complexity of the laws, has made the framework a perceived barrier to big data use. However, the framework permits many big data uses that are beneficial to public health.

Federal laws and regulations

The Health Information Portability and Accountability Act of 1996 (HIPAA). The HIPAA Privacy Rule governs "protected health information" (PHI), which is individually identifiable information about an individual's care, health condition, or payment for care.[9] The Privacy Rule does not govern "de-identified information,"[10] information from which 18 identifiers are removed or that an expert determines carries a minimal risk of being able to identify an individual if used.[11] The Privacy Rule applies to "covered entities" (i.e., health

plans, health-care clearinghouses, and most health-care providers[12]) and their "business associates" (i.e., entities having access to or using PHI when performing specified functions or services for the covered entity[13]), collectively referred to herein as "regulated entities."

Regulated entities are required to disclose PHI to the individual subject of the information or his/her designated representative and to the Secretary of the U.S. Department of Health and Human Services for purposes of investigations or enforcement. Regulated entities may disclose most PHI to anyone else with the individual's authorization, and are permitted (but not required) to disclose PHI without authorization in accordance with one of many permissive disclosure exceptions.[14]

Treatment, payment, and operations. Generally, regulated entities may disclose PHI without authorization for treatment purposes (e.g., patient care delivery), payment activities (e.g., payment for services), and health-care operations (e.g., quality improvement efforts).[15]

Public health activities. Regulated entities may disclose PHI without authorization to legally authorized public health entities for purposes related to preventing or controlling disease, injury, or disability (e.g., disease reporting, surveillance, and interventions).[16] Subject to limitations, the Privacy Rule identifies six other public health activities for which disclosure is permissible, including disclosure to a person exposed to a communicable disease and immunization reporting to a school.[17]

Other exceptions. The Privacy Rule identifies 11 other purposes for which disclosure may be made without authorization.[18] These exceptions are limited to specific activities that are beneficial to the public, including for national security purposes, certain law enforcement activities, and research, if a privacy or an institutional review board (IRB) has waived or altered the authorization requirement, or if the PHI will only be used for certain limited purposes.

Limited datasets. Regulated entities may disclose a limited dataset without authorization for research, public health, or health-care operations.[19] A limited dataset is PHI devoid of 16 specified identifiers, but may include city, state, and ZIP code; dates; and characters or codes that are not direct identifiers. The parties exchanging the dataset must enter into a data use agreement governing the use of the limited dataset(s).

The Common Rule. The Common Rule protects most human subjects involved in federally funded research[20] as well as individually identifiable private information obtained from a subject,[21] and generally requires either IRB approval and patient consent or IRB waiver of the consent requirement.[22] The Common Rule does not govern studies conducted using existing patient information, observation of public behavior, or survey or interview procedures. The Common Rule also exempts studies using existing data, records, or bio-specimens if the results do not reveal the subject's identity or if the data sources are publicly available.[20]

The Genetic Information Nondisclosure Act of 2008 (GINA). GINA generally prohibits health plans from using genetic information to make coverage-related decisions and requesting that beneficiaries undergo genetic testing or provide genetic information.[23] GINA also generally prohibits employers from discriminating against employees or applicants based on genetic information and from using genetic information in employment decisions, subject to exceptions.[24] Employers may disclose genetic information in certain circumstances, including to an occupational or health researcher and to a public health organization if the information relates to a contagious disease presenting an imminent threat of serious harm or death.[25]

42 C.F.R. Part 2 (Part 2). Part 2 applies to substance abuse programs that are federally assisted,[26] which includes providers who participate in Medicare, have a U.S. Drug Enforcement Administration number, or are federally tax exempt.[27] Providers must obtain written patient consent to disclose information that could identify that individual as a substance abuser,[28] with limited exceptions, including identifying disclosures for purposes of a state-mandated inquiry into a patient's cause of death,[27] child abuse reporting,[29] and certain research activities.[30]

State laws and regulations

States define their own privacy framework, which typically includes laws governing the same entities, activities, and/or types of information as the all federal laws.[31] Generally, providers must comply with all federal laws and any state requirements that are more protective.[32] States often provide enhanced protections for certain sensitive information (e.g., human immunodeficiency virus/acquired immunodeficiency syndrome test results[33] and mental health information[34]), and for some vulnerable populations (e.g., minors). State laws and regulations are relevant to the extent that they restrict disclosure of identifiable information more than federal laws, and entities that disclose or use identifiable information must be aware of how their state regulates sensitive information. Additionally, states

often require reporting of certain information related to communicable diseases, and generally maintain and mandate reporting to condition-specific (e.g., cancer) registries that support public health surveillance and research activities.

IMPLICATIONS FOR PUBLIC HEALTH PRACTICE AND POLICY

Understanding the federal framework for privacy is critical for big data use. While every law permits disclosure of health information with patient consent, obtaining consent may be impracticable, particularly for population-based activities. A number of exceptions permit disclosure of health information without consent, enabling public health functions. Unregulated domains outside the framework also present opportunities to make robust use of big data solutions.

Public health exceptions

Each federal law has carved out exceptions for basic public health services. HIPAA's exception is the broadest, permitting regulated entities to disclose PHI without authorization for activities related to preventing or controlling disease, injury, or disability.[16] Traditional public health activities such as tracking disease outbreaks, monitoring use of certain drugs, and targeting preventive screenings are all exempted. One recent example is a New York database created by public health officials, which tracks prescription drug disbursement so that providers can determine if an individual has an existing prescription.[35] In its first year, the database received seven million queries from 66,000 providers, reducing doctor shopping by 75%.

Health-care operations

HIPAA's exception for health-care operations is quite permissive, encompassing many population health activities. For example, the National Drug Early Warning System, created by the University of Maryland's Center for Substance Abuse Research,[36] monitors social media and traditional data sources. Using big data analytics, it detects emerging drug trends so that public health officials can launch community interventions to prevent the spread of illicit drug use.

Quality improvement. PHI may be disclosed for quality improvement activities, such as those facilitated by Minnesota's Reducing Avoidable Readmissions Effectively (RARE) project.[37] RARE uses hospital claims data to flag potentially preventable readmissions. Hospitals use this information to develop quality improvement interventions and redesign care processes. The RARE project has prevented nearly 8,000 readmissions.

Patient safety. Following a U.S. Food and Drug Administration (FDA) codeine prescribing alert, Navy and Marine Corps Public Health Center health analysis staff used health system repository data and big data solutions to analyze prescription rates and determine the prevalence of codeine prescribing practices.[38] The staff determined that providers had not modified their prescribing practices, and subsequently launched an outreach program to share information about the risks identified in the FDA alert, reducing codeine prescriptions in the relevant population by 99% in two months. This intervention is a patient safety activity, permissible under the HIPAA health-care operations exception.

Improving population health. PHI can be disclosed for population-based activities related to improving health or reducing costs. Population health management requires patient stratification, or identifying patients who will benefit from targeted interventions.[39] In 2001, Kaiser Permanente Northern California (KPNC) implemented a program to improve a 43.6% hypertension control rate. The program uses data from a central registry to stratify patients, automatically scan that list for gaps in care, and alert local practices when a patient's hypertension is not controlled. By 2009, hypertension control within KNPC reached 80.4%.[40]

Successful population health management relies on combining demographic, behavioral, and clinical data to develop more effective interventions. At Duke University, researchers integrated census data, county tax-parcel information, crime and housing statistics, and environmental data to support public health projects.[41] One county health department used the database to identify homes at high risk for childhood lead exposure, enabling the targeting of neighborhoods for interventions.

Outside the framework: unregulated domains

The legal framework for information privacy does not govern information that is de-identified, patient-generated, or in a nonregulated entity (e.g., a pharmaceutical company)'s possession. The following data sources are examples of data not subject to the health information privacy laws.

De-identified data. De-identified information can support health-related activities such as population-based investigations and research, the findings from which can be used to improve health-care quality and delivery. For example, IBM and Belgian pharmaceutical firm UCB collaborated in 2012 to improve epilepsy care.[42]

The team is processing 1.5 million epilepsy patients' de-identified data, which it will use to develop predictive analytics for use by a physician at the point of care to make treatment recommendations to the patient.

Patient-generated data. Individuals generate health information in myriad ways outside traditional health-care settings. Search engine queries about symptoms, over-the-counter drug purchases made with a pharmacy loyalty card, and social media data are all rich sources of information about an individual's health that can be used to perform public health functions. Recent examples include a Johns Hopkins program that can accurately predict where and when a flu outbreak will occur based only on tweets.[43] At Boston Children's Hospital, researchers can predict, track, and map obesity rates at the neighborhood level using only Facebook "likes."[44]

Non-regulated entities. Websites such as PatientsLikeMe.com collect and aggregate health information about individuals for sale and certain uses, such as medical research. When collecting information owned by non-regulated entities, users should be aware that general privacy laws and regulations may apply depending on the state.

CONCLUSION

As illustrated, the legal framework governing health information does not impede or prohibit many big data uses that support improvements in public health. Big data technologies can collect, store, and process population data, so that they can be analyzed and shared with stakeholders or de-identified for research and other purposes. Analytics solutions can evaluate population data in real time to stratify patient cohorts, identify high-risk individuals and populations, and alert authorities of potential outbreaks. Using big data solutions, public health authorities can work faster and more efficiently to develop and share knowledge that will improve the public's health. The legal framework for information privacy does not limit these possibilities, but rather facilitates them.

Jane Hyatt Thorpe and Elizabeth Gray are funded by the Robert Wood Johnson Foundation for work to develop and maintain an online resource of federal and state laws related to health information, including analyses, decision-support tools, and comparative maps (www.healthinfolaw.org). In addition, Thorpe is funded under a subcontract with ResDAC to provide guidance related to the Centers for Medicare & Medicaid Services' data-release policies. Thorpe also serves as a senior advisor in the U.S. Department of Health and Human Services Office of the National Coordinator for Health Information Technology (ONC). The authors thank Dr. Jeffrey Lerner and the ECRI Institute for addressing big data at their November 2013 annual conference and inviting Thorpe to speak, which prompted this article.

Jane Hyatt Thorpe is an Associate Professor and Elizabeth Gray is a Senior Research Associate in the Department of Health Policy at The George Washington University Milken Institute School of Public Health in Washington, DC.

Address correspondence to: Jane Hyatt Thorpe, JD, The George Washington University Milken Institute School of Public Health, Department of Health Policy, 950 New Hampshire Ave. NW, 6th Fl., Washington, DC 20052; tel. 202-994-4183; e-mail <jthorpe@gwu.edu>.

REFERENCES

1. Centers for Disease Control and Prevention (US), Office for State, Tribal, Local and Territorial Support. The 10 essential public health services: an overview. 2014 [cited 2014 Aug 3]. Available from: URL: http://www.cdc.gov/nphpsp/documents/essential-phs.pdf
2. Dumbill E. What is big data?: an introduction to the big data landscape. San Francisco: O'Reilly Media, Inc.; 2012 Jan 11. Also available from: URL: http://radar.oreilly.com/2012/01/what-is-big-data.html [cited 2014 Jul 22].
3. Mayer-Schönberger V, Cukier K. Big data: a revolution that will transform how we live, work, and think. New York: Eamon Dolan/Houghton Mifflin Harcourt; 2013.
4. Pearson JR, Brownstein CA, Brownstein JS. Potential for electronic health records and online social networking to redefine medical research. Clin Chem 2011;57:196-204.
5. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13:395-405.
6. van Panhuis WG, Grefenstette J, Jung SY, Chok NS, Cross A, Eng H, et al. Contagious diseases in the United States from 1888 to the present. N Engl J Med 2013;369:2152-8.
7. Kayyali B, Knott D, Van Kuiken S. The big-data revolution in US health care: accelerating value and innovation. Insights and Publications April 2013 [cited 2014 Jul 22]. Available from: URL: http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care
8. Edwards JE, Halawi LA. Analysis of variation in HIT privacy & security laws. Bus Rev Cambridge 2011;18:240-8.
9. 45 C.F.R. §160.103 at "protected health information" (2013).
10. 45 C.F.R. §164.502(d)(2) (2013).
11. 45 C.F.R. §164.514(b) (2013).
12. 45 C.F.R. §160.103 at "covered entity" (2013).
13. 45 C.F.R. §160.103 at "business associate" (2013).
14. 45 C.F.R. §164.502(a) (2013).
15. 45 C.F.R. §164.506 (2013).
16. 45 C.F.R. §164.512(b)(1)(i) (2013).
17. 45 C.F.R. §164.512(b)(1)(ii)–(vi) (2013).
18. 45 C.F.R. §164.512(a)(c)–(l) (2013).
19. 45 C.F.R. §164.514(e) (2013).
20. 45 C.F.R. §46.101 (2013).
21. 45 C.F.R. §46.102 (2013).
22. 45 C.F.R. §46.116 (2013).
23. GINA Title I §§101–104 (2008).
24. GINA Title II §§202–205 (2008).
25. GINA Title II §206(b) (2008).
26. 42 C.F.R. §2.12(e)(2) (2008).
27. 42 C.F.R. §2.15(b) (2008).
28. 42 C.F.R. §2.12(a)(1) (2013).
29. 442 C.F.R. §2.12(c)(6) (2013).
30. 42 C.F.R. §2.52 (2013).
31. George Washington University Hirsh Health Law and Policy Program. Health information & the law. Princeton (NJ): Robert Wood Johnson Foundation; 2012. Also available from: URL: www.healthinfolaw.org [cited 2014 Jul 24].
32. Pritts J, Lewis S, Jacobson R, Lucia K, Kayne K. Privacy and security solutions for interoperable health information exchange: report on state law requirements for patient permission to disclose health

information. Contract No. 290-05-0015. Washington: Office of the National Coordinator for Health IT (US); 2009.
33. Centers for Disease Control and Prevention (US). State HIV laws [cited 2014 Jul 24]. Available from: URL: http://www.cdc.gov/hiv/policies/law/states
34. Jost TS. Improving the quality of health care for mental and substance-use conditions: quality chasm series. Washington: National Academies Press; 2005.
35. Attorney General Eric T. Schneiderman. A.G. Schneiderman applauds success of New York's innovative program to prevent prescription drug abuse [press release] 2014 Feb 3 [cited 2014 Jul 26]. Available from: URL: http://www.ag.ny.gov/press-release/ag-schneiderman-applauds-success-new-york%E2%80%99s-innovative-program-prevent-prescription
36. National Institute on Drug Abuse (US). NIH system to monitor emerging drug trends [press release] 2014 Jul 17 [cited 2014 Jul 30]. Available from: URL: http://www.drugabuse.gov/news-events/news-releases/2014/07/nih-system-to-monitor-emerging-drug-trends
37. Institute for Clinical Systems Improvement, Minnesota Hospital Association, and Stratis Health. RARE: Reducing Avoidable Readmissions Effectively [cited 2014 Jul 27]. Available from: URL: http://www.rarereadmissions.org/index.html
38. United States Navy. Navy and Marine Corps Public Health Center helps reduce dangerous prescriptions in pediatric patients [press release] 2013 Aug 6 [cited 2014 Jul 27]. Available from: URL: http://www.navy.mil/submit/display.asp?story_id=75800
39. Gardner E. First steps for population health. Health Data Manag 2014;22:32-6.
40. Jaffe MG, Lee GA, Young JD, Sidney S, Go AS. Improved blood pressure control associated with a large-scale hypertension program. JAMA 2013;310:699-705.
41. Miranda ML, Ferranti J, Strauss B, Neelon B, Califf RM. Geographic health information systems: a platform to support the "triple aim". Health Aff (Millwood) 2013;32:1608-15.
42. IBM. UCB and IBM collaborate to personalize care for epilepsy patients [press release] 2013 May 16 [cited 2014 Jul 26]. Available from: URL: http://www-03.ibm.com/press/us/en/pressrelease/41083.wss
43. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic. PLoS One 2013;8:e83672.
44. Chunara R, Bouton L, Ayers JW, Brownstein JS. Assessing the online social environment for surveillance of obesity prevalence. PLoS One 2013;8:e61373.