KEEPING THE INTERNET
OPEN • INNOVATIVE • FREE
www.cdt.org
CENTER FOR DEMOCRACY
& TECHNOLOGY

Health Big Data in the Government Context

Data-driven and information-based systems have quickly become the new paradigm for American health care. Using the vast amounts of data produced from both traditional healthcare records and newer commercial sources like mobile health apps, the techniques of big data analysis can contribute significantly to the continued transformation of the health care landscape. Health big data can help clinicians to make more cost-effective, high-quality decisions, improve medical research, and more fully engage consumers in managing of their health.

Government agencies, which collect huge amounts of health data, are eager to take advantage of these opportunities. They are applying big data techniques internally, and they are providing unprecedented access to their health data holdings both by granting researchers access, and by opening data sets to the public.[1] However, the serious privacy and security considerations that arise from collection and use of sensitive health information pose a barrier to the realization of big data's benefits. Currently, health data privacy and security are addressed in a multitude of state and federal laws and regulations, which, in their very complexity, can seem to fail to offer comprehensive guidance on the ethical and responsible use of personal health information.

To explore the privacy and security implications of health big data,[2] and to develop concrete proposals for how to address those issues and at the same time reap the benefits of big data, CDT is undertaking a series of consultations with stakeholders and experts. We are examining three scenarios: (1) clinical and administrative data generated by health care providers and payers; (2) health data contributed by consumers using the Internet and other consumer-facing technologies; and (3) health data collected by federal, state, and local governments.

In this paper, we focus on the third of these scenarios: health data collected by federal, state, and local governments. We examine the current legal landscape for the collection and use of health data by governments. We ask provocative questions, such as, for what

[1] *See e.g.* CMS.GOV, Medicare Provider Utilization and Payment Data: Physician and Other Supplier, http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html (last visited Feb. 27, 2015) (payments to doctors

[2] We use "big data" as shorthand for developments in data collection and processing, characterized by rapid increases in the volume, velocity, and variety of data generated by information technologies. *See* Jane Sarasohn-Kahn, Here's Looking at You: How Personal Health Information Is Being Tracked and Used, California HealthCare Foundation (July 2014), http://www.chcf.org/publications/2014/07/heres-looking-personal-health-info; *See also* Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* (May 2014), http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf. For our purposes, the big data phenomenon encompasses not only the proliferation of "always on" sensing devices that collect ever larger volumes of data but also the rapid improvements in processing capabilities that make it possible to easily share and aggregate data from disparate sources and, most importantly, to analyze it and draw knowledge from it.

secondary purposes should government health data be used or disclosed? Most importantly, we seek to identify ways in which the collection and use of health data by governments can be performed in privacy and security-protecting ways. As our guide, we use the Fair Information Practice Principles (FIPPs), a framework that has informed most modern data privacy laws.

Governments As Health Data Stewards

Governments in the United States collect and use health data in multiple ways for a wide range of purposes and functions:

- State and local governments are direct providers of medical services. There are over 1,000 publicly owned, non-federal hospitals in the US, which in 2008 represented 22% of all hospitals and provided care for 14% of all inpatients. State and local agencies collect health data as they provide a wide variety of other services, such as mental health services, drug and alcohol treatment, rural health initiatives, and programs that address healthcare disparities.[3]
- The federal government too is a direct provider of health care services. The Veterans Administration health system, one of the largest providers in the country, has an enrolled population of over 9 million and provides health care to over 6.5 million patients a year.[4]
- Governmental entities pay for a very large share of health care in the US, mainly thorough Medicare and Medicaid. In 2010, Medicaid accounted for 16% of health care coverage and Medicaid 14.5%.[5] By 2023, health expenditures financed by federal, state, and local governments are projected to account for 48 percent of national health spending.[6]
- To support health care and payment reform initiatives, a growing number of states are establishing all-payer claims databases (APCDs). The information typically collected in an APCD includes patient demographics, provider codes, and clinical, financial, and utilization data. States are also creating or supporting health information exchanges to facilitate exchange of electronic health records among providers.[7]
- Governments collect or access health data for regulatory and safety purposes. For example, 49 states maintain prescription drug databases, intended to both identify people who acquire excess prescriptions for addictive drugs like

[3] Taressa Fraze et al., *Public Hospitals in the United States, 2008* (Sept. 2010), http://www.hcup-us.ahrq.gov/reports/statbriefs/sb95.pdf.

[4] Erin Bagalman, *The Number of Veterans That Use VA Health Care Services: A Fact Sheet* (June 3, 2014), http://www.fas.org/sgp/crs/misc/R43579.pdf.

[5] HEALTHPACONLINE.NET, Healthcare Statistics in the United States http://www.healthpaconline.net/health-care-statistics-in-the-united-states.htm (last visited Feb. 27, 2015).

[6] Centers for Medicare & Medicaid Services, *National Health Expenditure Projections 2013-2033,* http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/Proj2013.pdf.

[7] Joe Porter et al., *The Basics of All-Payer Claims Databases: A Primer for States* (Jan. 2014), http://www.rwjf.org/content/dam/farm/reports/issue_briefs/2014/rwjf409988.

painkillers and tranquilizers, and the physicians who overprescribe them.[8]

- Governments collect and use personal data in the fulfillment of their public health role, a practice called "public health surveillance."[9] Public health data collection focuses not only on individual biological and behavioral characteristics, but also on social and environmental factors.[10] Many states require providers to report certain diseases and conditions to the state health department for entry into disease-specific registries. These registries identify and track patients with a specific diagnosis, typically based on geographic location or treatment hospital, and often include information collected voluntarily from patients. New York State's Department of Health, for example, maintains disease registries for chronic and communicable diseases such as cancer and HIV-AIDS.
- The United States government also collects health information through other activities. For example, the Internet Revenue Service collects information about medical tax deductions through routine filings, and may investigate in more detail in the event that fraud is suspected. The Environmental Protection Agency collects information about how pollutants may be detrimental to individuals; this information is not always personal, but it is most often used to make decisions that have a significant impact on both individuals and public health. Government agencies also collect a wide range of data that have relevance to health care, such as vital statistics data recording the time, day, and location of our births, and the date and cause of death.

In these and other instances, across a range of functions, government health officials are able today to gather and use more data about more individuals and communities, combined with more sophisticated analysis tools. The trends will undoubtedly continue. The response to the recent Ebola outbreaks in Africa, for example, shows how officials can mobilize an array of technological and analytic tools in response to a health crisis. The National Geospatial Intelligence Agency released a mapping application that provides real-time information on outbreaks, along with the location of medical treatment centers, electrical grids, and water supplies. An open source application deployed by the Centers for Disease Control (CDC) identifies and catalogs people exposed to the virus, along with symptoms and other patient history, to help those on the ground recognize patterns and allow them to predict the next outbreak. Another example of innovative applications of health information by the government comes from the Food and Drug Administration (FDA). The FDA, under Congressional mandate, is building a database system called Sentinel to monitor the safety of FDA-approved drugs and medical

[8] Alan Schwarz, *Missouri Alone in Resisting Prescription Drug Database,* N.Y. TIMES (July 20, 2014), http://www.nytimes.com/2014/07/21/us/missouri-alone-in-resisting-prescription-drug-database.html.

[9] *See* U.S. Dept. of Health and Human Services, *CDC's Vision for Public Health Surveillance in the 21st Century* (July 2012), http://www.cdc.gov/mmwr/pdf/other/su6103.pdf ("Public health surveillance is the systematic, ongoing collection, management, analysis, and interpretation of data followed by the dissemination of these data to public health programs to stimulate public health action. The best recognized use of public health surveillance data is the detection of epidemics and other health problems in a community.").

[10] Gerald Midgley, *Systemic Intervention for Public Health*, 96 Am. J. Public Health 466 (2006), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1470519/.

products after they reach the market.[11] The system was built and is being deployed in a privacy protective manner by using a distributed network and asking data sources to apply specific questions to data they are already collecting and then sending the answers to the FDA.

The Legal and Policy Landscape

The sensitive nature of identifiable health data, along with the possibility of it being used to make inferences about individuals, amplifies the government's responsibility to adhere to stringent data collection and access controls, reasonable use limitations, security measures, and reasonable deletion and de-identification practices. However, the laws surrounding the collection and use of health data by government agencies are complex and multi-layered, in part because the role of governments in health care is complex and multi-layered.

In general, the sharing of non-public health data by governments is delineated by the distinction between identifiable and non-identifiable data. Identifiable data is restricted in three general ways: by internal rules, regulatory requirements, and federal legislation. Most broadly applicable is the Privacy Rule adopted under the Health Insurance Portability and Accountability Act (HIPAA),[12] which regulates the use and disclosure of individually identifiable information by covered entities. The HIPAA Privacy Rule applies to many governmental providers and payers.[13] Other federal laws also regulate health data. For example, Section 308 (d) of the Public Health Service Act[14] limits the release of sensitive health data that is either identifiable or potentially identifiable for any unexpected purposes. Part 2, Title X of the same law requires healthcare providers to get written consent before they are allowed to disclose patient information to other organizations, including the government.

In additional to the federal government, most states have their own laws and regulations pertaining to the use, collection and disclosure of health information.[15] For example,

[11] *See* U.S. Dept. of Health and Human Services, *Access to Electronic Healthcare Data for More than 25 Million Lives: Achieving FDAAA Section 905 Goal One* (July 2010), http://www.fda.gov/downloads/Safety/FDAsSentinelInitiative/UCM233360.pdf.

[12] *See* HHS.GOV, Summary of the HIPAA Privacy Rule, http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/ (last visited Feb. 27, 2015).

[13] For example, the Centers for Medicare and Medicaid Services (CMS) is a HIPAA-covered entity, as is the Veterans Health Administration. *See* Dept. of Veterans Affairs, *Protected Health Information (PHI) and Busisness Associates Agreements Management* (Sept. 2014), http://www1.va.gov/vapubs/viewPublication.asp?Pub_ID=761&FType=2.

[14] 42 U.S.C. § 242(d)

[15] *See* Health Information Security and Privacy Collaboration, *Update to October 9, 2007 Report on State Privacy & Security Laws Related to Electronic Health Records and Electronic Health Information Exchange* (June 2008), http://www.healthit.gov/sites/default/files/hspl_1_leg_analysis508.pdf; *See also* Joy Pritts et al., Privacy and Security Solutions for Interoperable Health Information Exchange: Report on State Law Requirements for Patient Permission to Disclose Health Information (Aug. 2009), http://www.healthit.gov/sites/default/files/disclosure-report-1.pdf. For other resources on state law, *see* HEALTHIT.GOV, Health Information Security & Privacy Collaboration (HISPC), http://www.healthit.gov/policy-researchers-implementers/health-information-security-privacy-collaboration-hispc (last visited Feb. 27, 2015).

"nearly every state has some statutory or regulatory provisions that grant individuals the right to access their medical records maintained by medical doctors and/or hospitals."[16] HIPAA pre-empts state laws that are less privacy-protective, but allows states to impose more stringent requirements. Many states also have separate laws focused on providing protection for specific categories of sensitive information, such as disease diagnosis, mental health and substance abuse issues, and family planning activities. State laws also mandate minimum retention periods for health records.[17]

In addition to these laws that are specific to health data, federal government agencies are subject to the Privacy Act of 1974, which regulates federal agencies' collection, maintenance and distribution of personally identifiable information. The Privacy Act applies to information held in a "system of records" that uses identifying factors like names and social security numbers. Generally, the Act requires agencies to obtain an individual's consent before disclosing data, but there are twelve exceptions to the consent requirement for disclosure in the Act, including an exception for internal uses and for disclosures affecting the health or safety of an individual. In addition, through the "routine use" exception, agencies have broad discretion to use public data for general health purposes. Further, when information is de-identified, the Privacy Act does not cover it and government entities do not need a person's consent to collect and use it. The Privacy Act also includes access and correction mandates that give Americans some control over personal information — including health information — collected about them by federal agencies. It gives people the right to know what information was collected about them, to see and have a copy of that information, to correct or amend that information, and to exercise limited control over disclosure of that information to other parties. The law applies to all federal agencies, including those that collect health or health-relevant data. This means that any federal agency that provides healthcare services for the government, such as the Veterans Administration, would likely be subject to both HIPAA and Privacy Act regulations.[18]

All of these laws are based, to greater or lesser degrees, on the Fair Information Practice Principles (FIPPs), [19] a widely accepted framework for information privacy. Some argue

[16] Joy Pritts et al., Privacy and Security Solutions for Interoperable Health Information Exchange: Report on State Medical Record Access Laws (Aug. 2009), http://www.healthit.gov/sites/default/files/290-05-0015-state-law-access-report-1.pdf.

[17] HealthIT.gov, State Medical Records Laws: Minimum Medical Record Retention Periods for Records Held by Medical Doctors and Hospitals, http://www.healthit.gov/sites/default/files/appa7-1.pdf (last visited Feb. 27, 2015).

[18] 5 U.S.C. § 552a(m)(1).

[19] There is no definitive version of the FIPPs. We use an articulation of the FIPPs drawn from three sources: the Markle Connecting for Health Common Framework, the White House's 2012 Consumer Bill of Rights and the ONC's Nationwide Privacy and Security Framework for Electronic Exchange of Individually Identifiable Health Information. *See* Markle.org, MArkle Common Framework: Sharing and Protecting Information, http://www.markle.org/health/markle-common-framework (last visited Feb. 27, 2015); *See also* The White House, *Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy* (Feb. 2012), http://www.whitehouse.gov/sites/default/files/privacy-final.pdf; *See also* Office of the National Coordinator for Health Information Technology, *Nationwide Privacy and Security Framework for Electronic Exchange of Individually Identifiable Health Information* (Dec. 2008), http://www.healthit.gov/sites/default/files/nationwide-ps-framework-5.pdf. The three formulations incorporate the same principles, but they do not perfectly align, so we have combined them into a single framework.

that the FIPPs are unsuited for the era of big data. We believe, however, that the FIPPs framework has withstood the test of time and technology. The framework is flexible yet structured and informs most modern privacy regimes inside and outside healthcare. CDT believes the FIPPs offer governments seeking to use big data with regard to health information a strong, standardized structure that helps make sense of the confusing patchwork of state and federal laws in this arena, promoting responsible and efficient uses of data while allowing for innovations in analytics and application. The remainder of this paper analyzes individual FIPPs as they apply to contexts in the government collection, use, and disclosure of health information, and makes recommendations for governments that embark on these activities.

Openness / Transparency

Though some posit that today's opaque data collection and usage practices make the foundational principle of openness and transparency obsolete, we believe it remains completely possible to implement, and is crucial to any data regime. Transparency should guide any health data collection and use regime, from the first point of contact with data to subsequent uses. Regardless of how and when data is collected, detailed information about how specified health data sets are collected and used, even when de-identified, should be detailed in a statement accessible to the public in one place. Whether or not most citizens routinely access these detailed notices, they promote external accountability from regulators and interested stakeholders. Such notices also provide internal guidance to those who manage data, so that they understand the rules. For causal readers who are generally curious about data practices, it would be best to layer the detailed notice to provide top-level, digestible information. Health care providers should also consider whether there are lessons to be learned in terms of just-in-time delivery of notices, which has been explored in the context of mobile apps.[20] Just-in-time notices could provide information in more digestible and context-specific increments.

Governments should endeavor to provide more prominent notice to individuals whenever the intended usage or collection might be either unexpected or objectionable. Though it is reasonable to assume that most people interacting with their governments understand that collection and use of personal information by government is part of citizenship, it is also fair to assume that most people might not expect their governments to have detailed health data about them, unless they were enrolled in a government health program such as Medicaid or Medicare. Further, it is unclear how much people understand about the information collected for public health purposes. While there are clear, societally beneficial reasons for collecting and using this data, authorities should perhaps endeavor to find more ways to inform individuals. When non-governmental third parties collect and share data with the government, contextual notice – or, just-in-time notice – becomes a critical component of meeting an individual's expectations regarding collection and sharing. This is particularly true on government-sponsored websites, where third parties like advertising networks are often employed to provide web analytics

[20] *See* Federal Trade Commission, *Mobile Privacy Disclosures: Building Trust Through Transparency* (Feb. 2013), http://www.ftc.gov/sites/default/files/documents/reports/mobile-privacy-disclosures-building-trust-through-transparency-federal-trade-commission-staff-report/130201mobileprivacyreport.pdf.

or retargeting services. The Office of Management and Budget has provided federal agencies with clear guidance[21] on the elements of effective public notice when using third parties, such as how to distinguish branding and best practices for privacy policies. This guidance is a useful roadmap on transparency for state and local agencies as well.

Entities that provide health-related services to governments but do not directly interface with citizens should also clearly describe their data practices in privacy statements available on their websites. As with government privacy policies, few people will directly access that information, but making it available to regulators, advocates, employees and interested parties will incentivize such vendors to adopt objectively reasonable practices, and will serve as a basis for accountability to those policies. In order to be most effective, these notices should try to be as specific as possible about data collection and use practices, though it may be appropriate to rely on representative examples for certain data uses or sharing. Vendors should not seek to rely on broad and inscrutable reservations of rights that are difficult for even sophisticated readers to interpret.

Purpose Specification and Use Limitations / Respect for Context

Respect for context means that consumers have the right to assume that government entities will collect, use, and disclose personal data only in ways that are consistent with the context in which that data was provided. As we have noted in previous papers,[22] we believe that institutions holding personal health data should be able to use that data internally for secondary purposes, such as operational analytics and generalizable research, though deciding which research is justified and worthwhile requires that agencies perform in-depth ethical assessments. Information should be publicly available on these secondary uses, but no new consent should be required generally for operational analytics and generalizable research with existing data sets; requiring consent would unreasonably burden valid research and analytics efforts.[23] On the other hand, transfer of sensitive data by the government in identifiable form to third parties — even for research or other societally beneficial purposes — may run counter to ordinary individual expectations, and should only happen with informed consent.

We suggest these criteria should be considered in determining whether a secondary use going beyond operational analytics and generalizable research is consistent with the context in which the data was originally collected, and thus may not require new permission from the user: (1) where no extra data collection, retention, and transfer occur (apart from transfer to dedicated service providers with no independent right to use the data), (2) where the secondary use of the data is not used to materially alter an individual's experience, and (3) where the output is aggregate and not personally identifiable. Meeting these criteria will not necessarily be determinative, but should

[21] Executive Office of the President, *Memorandum for the Heads of Executive Departments and Agencies: Guidance for Agency Use of Third-Party Websites and Applications* (June 2010), http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-23.pdf.
[22] Center for Democracy & Technology, *Encouraging the Use of, and Rethinking Protections for De-Identified (and "Anonymized") Health Data* (June 2009), https://cdt.org/files/pdfs/20090625_deidentify.pdf.
[23] Our concept of operational analytics and generalizable research does not include marketing.

inform whether to provide more information to a person about secondary uses, or whether to affirmatively ask the person for permission.

A far more difficult context question arises when private service providers that hold identifiable personal health information turn this data over to governments. On the one hand, at least some personal information should be — and indeed is required by law to be — turned over to public health officials in order to protect the interests of the general public. For at least some of these transfers, consent will not be required, and individuals will also not be able to opt out of the data collection. On the other hand, we do not presume that government officials are entitled to *all* personal health information merely because it may one day be useful in treating or preventing the spread of disease. The public health community should explore with other stakeholders how to define acceptable limits on both mandatory reporting and voluntary sharing of data for public health purposes. The phenomena associated with big data — interoperability, ease of collection and sharing, and the ability to draw more knowledge from data — create incentives for governments to impose more extensive reporting requirements, and for providers to engage in more voluntary sharing. We recommend that public health authorities exercise caution in imposing more expansive reporting requirements. The public health exception to the use limitation principle should not be expanded without meaningful justification. To the extent possible, service providers who share personal information for public health reasons should try to highlight such sharing in a contextual notice that provides detailed information about what data is being collected, how it is being used, and with whom it is being shared.

Focused Collection / Collection Limitation

In the era of big data, governments, like commercial entities, have much greater technical capacity to collect data. This applies to all the various ways in which governments collect health-related data. Across all of the health-related functions of government — direct services, payment, public health, drug treatment, welfare, in-school programs, research — it is easier than ever to collect data. In its role as payer, the government has substantial power to collect data from providers; regulations stemming from CMS require providers to collect and share with the government; increasing data about a range of factors relevant to quality and cost, and big data analytic capabilities vest that data with potentially more value than ever before. Moreover, government entities collecting health data directly from citizens might positively augment that data with information from commercial sources. For example, government health providers could access health data generated by commercial gadgets and applications, such as Jawbone and Fitbit, through partnerships with these services. Government researchers may access data sets from non-governmental sources. Conceivably, governments could even inform decisions about benefit programs by accessing data collected by broker services on individuals' lifestyle choices and habits.

A fundamental principle of privacy is that an entity (governmental or private) should collect no more information than is necessary for the transaction at hand. The White House privacy report called this "focused collection." Even in the age of big data, and even with respect to governmental activities, focused collection remains a valid principle, reflected both in HIPAA's "minimum necessary" standard and in the federal Privacy Act.[24]

One way to describe the collection limitation principle is that governments should not collect information about individuals in ways that would surprise them, or are contrary to the interests of individuals.[25] In general, any entities collecting data should consider the average person's reasonable expectations. Some peripheral data collection, such as that collected outside of typical processing, may be reasonably necessary for claims processing and fraud prevention, but such collection should be focused and the individual should be notified — in advance or, if there is a legitimate rationale to delay, as soon as possible after the fact.

The principle of focused collection can allow extensive collection, especially in connection with government payment programs. A person applying for a government health benefit should expect that the government would collect any information that is relevant to eligibility determination, including information relevant to protecting the government from fraud. The government is responsible, of course, for fairly evaluating all of this data as a due process matter, and for protecting its security.

Minimization and De-identification

The principle of data minimization holds that entities should collect no more information than is necessary for the stated purpose, and should retain it no longer than is necessary to serve that purpose. In the health context, governments (and researchers using government data) should not collect any and all data points in the hopes that this data may someday be useful. Over-collection of identified data and indefinite retention of data pose risks, even when the data was collected for benign reasons. These risks include data breach, internal misuse, improper disclosure or secondary use, and government use for non-health-related purposes.[26] Ignoring the minimization principle could also have a "chilling effect" on individual behavior: patients, especially those in the most vulnerable populations, might be less inclined to seek treatment and less inclined to fully share critical information if they perceive that government health care providers and others are collecting information unnecessarily and without attention to its potential misuse. This effect may be particularly acute in the health data context: recent research

[24] The Privacy Act specifies that federal agencies may maintain in their databases "only such information about an individual as is relevant and necessary to accomplish a purpose of the agency required to be accomplished by statute or by executive order of the President." 5 U.S.C. 552a(e)(1). Strictly speaking, HIPAA's minimum necessary standard applies "[w]hen using or disclosing protected health information or when requesting protected health information from another covered entity or business associate." 45 CFR 165.502(b). Nevertheless, a comprehensive privacy framework limits collection as well as use and disclosure.

[25] In August 2010, the HHS Health IT Policy Committee adopted as a core value the principle that "Patients should not be surprised about … by collections, uses or disclosures of their information."

[26] Justin Brookman and G.S. Hans, *Why Collection Matters: Surveillance As A DeFacto Privacy Harm, available at* http://www.futureofprivacy.org/wp-content/uploads/Brookman-Why-Collection-Matters.pdf.

has found that American consumers consider health data the second most sensitive data type (after social security numbers).[27]

However, it is argued that data minimization is antithetical to the goals and potential of big data. If entities are required to minimize collection and delete data when no longer necessary for the purpose for which it was collected, the thinking goes, they will be unable to realize the upsides of data analytics – especially future research not contemplated at the time the data was collected.

One important method in minimization is de-identification, combined with enforceable prohibitions against re-identification.[28] Many of the public health, quality control, and research benefits of big data analysis can be achieved with de-identified data, though this use standard does not, of course, provide cover for a free-for-all of data collection with no limitations. Indeed, the governmental interests in public health, health research, and management of health care costs offer a powerful justification for de-identification, and for ongoing research into methodologies that minimize risk. De-identification should be the default in circumstances where the government releases data, whether in a closed or public way. Public release requires de-identification to a level necessary to control the risk of re-identification where other controls (such as contractual controls) are not available or cannot be relied on. Given the push towards open data, agencies need to be especially cognizant of the risk that seemingly unrelated data sets – containing location, health, and demographic data – may be combined to result in surprisingly detailed aggregate individual records. While prohibiting re-identification of de-identified data may limit risks inside the perimeter of an agency and while contracts can be used to bind downstream recipients, public release of data is the most hostile environment for big data with respect to re-identification threats, and government entities must do it especially well. For example, veterans' health records have a higher risk for being re-identified (after de-identification) than other government-held records, largely because veterans are a smaller population. To reduce this risk, the VA generally releases small data sets, and only investigators with a VA affiliation (rather than entities outside of the VA), in order to meet data minimization and collection limitations standards.

Any data that governmental agencies release should be examined prior to public distribution in order to determine what, if any, health data is contained within the set, how to remove identifying information in order to protect the privacy rights of individuals, and how the data might be combined with other data sets in unanticipated ways.

The de-identification debate and the public release of large government data sets are inextricably intertwined. In order for governmental agencies to release those sets, they will need to determine if any de-identification procedures should take place beforehand, and what the risks of different de-identification strategies may be. They can then implement those procedures with a strong level of certainty that they will actually be

[27] Pew Research Center, Public Perceptions of Privacy and Security in the Post-Snowden Era (Nov. 2014), http://www.pewinternet.org/2014/11/12/public-privacy-perceptions/.

[28] Government agencies should carefully review their de-identification practices and should collaborate with researchers to keep abreast of de-identification techniques and the risks of re-identification, to be sure that the government is using the most advanced techniques.

effective. The limited data set, which requires fewer identifier categories to be removed than is necessary to qualify as "de-identified," offers one important vehicle for government uses without needing to obtain authorization. The recipient of the limited data set must sign a data use agreement that sets forth the purpose for which the data can be used, and which prohibits re-identification.

Outside of a limited data set, government might consider establishing two levels of de-identification for release of data: 1) total de-identification, which would allow for publication of the data, and 2) reasonably de-identified data, released only on the condition that recipients will not share it.

While strong de-identification is not a silver bullet that automatically eliminates any risk of re-identification or obligation to limit collection, there *are* methods that governmental entities can implement in order to largely (if not completely) reduce that risk, especially if de-identification is seen as a dynamic process. In this context, the perfect is not the enemy of the good.

<u>Ultimately, reasonable de-identification can support the creation of usable data sets for public research</u>. Already, government agencies have shown that is possible to perform strong de-identification and apply this data toward civic projects. Healthdata.gov, for example, is an online federal resource of de-identified public health information. The site includes a "hospital compare" tool that details process of care, mortality of patients, and readmission quality measures, as well as TOXMAP, a tool that helps users create maps of chemical contaminants identified in geographic areas. The Affordable Care Act (ACA) authorizes HHS to release new data resources that advance transparency, accuracy and access in the health care provider market and health insurance market in significant ways. HHS has begun "liberating" health data through the Health Data Initiative, making more and more data from HHS' vaults (from CMS, CDC, FDA and NIH, to name a few sources) easily available and accessible to the public and to innovators across the country.

This information includes details about the quality of clinical care providers, nationwide health service provider directories, databases of the latest medical and scientific knowledge, consumer product data, community health performance information, government spending data and much more. Other programs are using de-identified or anonymized private sector data for governmental health purposes. For example, in order to target unreported outbreaks of foodborne illness, New York City's Department of Health and Mental Hygiene analyzed nine months of de-identified data from Yelp reviews that included words like "vomit" and "food poisoning." Health inspectors were subsequently deployed to restaurants with the most frequent complaints, and found health code violations at each place.

With ongoing improvements in the understanding – and techniques – of de-identification, it should be possible 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."[29]

Data Integrity and Quality / Data Access and Accuracy

Closely associated with the principle of data accuracy is the principle of access: individuals should be able to access and copy data about themselves. Accuracy of, and access to one's own personal data, should be critical components of government data collection and use programs. HHS's Blue Button Initiative is successfully giving patients' access to their own data. First launched by the Department of Veterans Affairs, and now embraced by both the Centers for Medicare and Medicaid Services and major private health plans, the Blue Button is a public-private partnership that gives consumers easy and secure access to their health records from a variety of sources. An individual's records are available in formats they can use in conjunction with applications and services that help individuals to analyze and manage their health data.

Depending on the circumstances, government entities may also have a legal obligation to provide individuals with access to their health records and the ability to amend these records. The HIPAA Privacy Rule applies to government-sponsored health programs such as military, veteran, or government employee plans. Therefore participants in these plans will always have the right to access and inspect health records maintained by these programs. Participants may also request that a health plan correct a record. However if the plan denies the correction request, the participant may only submit a letter of disagreement to be added to their record.[30] In addition to potential HIPAA regulation, all federal agencies are subject to the 1974 Privacy Act, which regulates the collection, maintenance and distribution of citizens' personally identifiable information. Although medical providers are only required to share patient information with the government in prescribed circumstances (such as for public health reporting or if under subpoena, though there is discretion for these categories to grow), and HHS does not maintain a government database of citizens' health records unless they are Medicare beneficiaries, HHS may obtain information on a patient during privacy and security compliance audits or when investigating HIPAA violations.[31] In this and other ways, HHS could compile a sizeable record on a persons' health history that would be subject to the Privacy Act. CMS – which collects and stores data from Medicare and Medicaid recipients – would likely be subject to the Privacy Act as well. Therefore, any U.S. citizen or permanent resident would have the right to access, inspect and request amendment

[29] Deven McGraw, *Building Public Trust in Uses of Health Insurance Portability and Accountability Act de-identified data*, J Am Med Inform Assoc 1 (2012), https://cdt.org/files/pdfs/JAMIA-de-identified-data-trust.pdf.
[30] HHS.GOV, Your Medical Records, http://www.hhs.gov/ocr/privacy/hipaa/understanding/consumers/medicalrecords.html (last visited Feb. 27, 2015).
[31] HHS.GOV, Does the HIPAA Privacy Rule Require My Doctor to Send My Medical Records to the Government?, http://www.hhs.gov/ocr/privacy/hipaa/faq/disclosures_for_rule_enforcement/347.html (last visited Feb. 27, 2015).

of HHS or CMS health records.[32] These individuals may also appeal a denial of their request to amend to an administrative authority.[33]

It should be noted that the Privacy Act does not cover non-citizens, even though there are many groups of non-citizens who qualify for U.S. healthcare coverage, and the Department of Health and Human Services has said it will give administrative rights to non-citizens who request their records.[34] It should be noted that the Privacy Act does not cover non-citizens, even though there are many groups of non-citizens who qualify for U.S. healthcare coverage. [35] Government-sponsored health programs such as Medicaid, Refugee Medical Assistance, and F1 (student) visa insurance, for instance, cover thousands of individuals in the United States each year.[36] If participant health data is collected through such programs, the Privacy Act would not protect these individuals' data, unless HHS adopted a policy to treat this data as a "system of records" subject to the Act's regulations.[37] These groups *may* be able to access, inspect, and request to change their health record through HIPAA if they are on HIPAA-covered health plans like Medicaid; however, the penalties and amendment procedures for HIPAA are arguably less stringent nor is there any right to a formal appeal of a denial to amend.[38] Thus HIPAA may not sufficiently address these groups' privacy concerns. This is especially problematic because a large portion of individuals participating in these programs is seeking asylum, refugees, trafficking victims or domestic violence survivors whose health records are highly sensitive.

Individual Participation / Control

A thoughtful application of the FIPPs should include consideration of whether policies enhance transparency to individuals about big data uses, enable more active engagement and input from individuals in research as well as access to research results, and include strong security protocols to protect against risks of internal misuse (such as unauthorized access) or inadvertent exposure of such data. Given the limitations of current law, controls must be put in place to ensure government collection and sharing of individual data is reasonably scoped. For unexpected data collection that could be objectionable to an ordinary person, governments should endeavor to get affirmative

[32] 45 C.F.R. § 5b.6 (2007) *available at* http://www.gpo.gov/fdsys/pkg/CFR-2007-title45-vol1/pdf/CFR-2007-title45-vol1-sec5b-6.pdf.
[33] 45 C.F.R. § 5b.7 (2007) *available at* http://www.gpo.gov/fdsys/pkg/CFR-2007-title45-vol1/pdf/CFR-2007-title45-vol1-sec5b-7.pdf.
[34] HEALTHCARE.GOV, Immigration Status and the Marketplace, https://www.healthcare.gov/immigrants/immigration-status/ (last visited Feb. 27, 2015).
[35] *Id.*
[36] *See* Dept. of Health and Human Services, *Services Available to Victims of Human Trafficking: A Resource Guide for Social Services Providers* (May 2012), http://www.acf.hhs.gov/sites/default/files/orr/traffickingservices_0.pdf.
[37] The Department of Homeland Security adopted a similar policy in 2007. *See* U.S. Dept. of Homeland Security, *Privacy Policy Guidance Memorandum: DHS Privacy Policy Regarding Collection, Use, Retention and Dissemination of Information on Non-U.S. Persons* (Jan. 2007), http://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2007-1.pdf.
[38] HHS.GOV, Your Medical Records, http://www.hhs.gov/ocr/privacy/hipaa/understanding/consumers/medicalrecords.html (last visited Feb. 27, 2015).

permission prior to collection. Data minimization — and particularly de-identified data sets — should also be used where possible to reduce the risks of hacks, internal misuse, unexpected secondary use, and overbroad government access. De-identifying data is especially important when data is being presented publicly, unless the person has consented to the use of their identifiable data. Even with such protections, there is still a possibility that misuse can occur with de-identified data. For example, if a subpopulation is denied benefits or treated in a discriminatory way using de-identified data, this has a negative impact on individuals even without re-identifying or targeting them specifically. Thus, additional user controls may need to be implemented at the point of collection to enhance consumer protections. One solution may be to require governments to provide detailed, meaningful notice to individuals on how identifiable and de-identified data sets will be used in the future, and to allow individuals to opt out of certain uses. In those instances where obtaining consent and/or allowing opt-out would unreasonably burden research and analytics efforts, the government should, at the very least, provide notice and information on future data use, and be certain that this standard is not being applied in an overly broad way.

Security

Effective security standards for data sets currently exist; the challenge is ensuring that governmental entities employ those standards in order to protect records. Governmental entities, like private actors, must take into consideration: a) the sensitivity of their records; b) the likelihood of breach or misuse; and c) currently effective security measures. The use of internal auditing and oversight to ensure that security methods are followed and up-to-date is essential for governmental entities to protect individual privacy and data records.

De-identification, as discussed above, is one step that can help protect the security of data records. Encryption and limitations on sharing are other clear actions that governmental entities can use to promote privacy protections. Consistently assessing the current state of risk, and potential vulnerabilities, will help promote reasonable and responsible security measures.

Administrative and regulatory requirements of federal, state, and local governments can limit data sharing. Security concerns and regulations, multimode displays (e.g., displaying data both in hard copy and web-based formats), and required use of specific software for data dissemination can affect timeliness and the ability to release data. These requirements can secure data and computer systems, and ensure patient, enrollee, or respondent privacy and confidentiality. However, substantial programmatic resources and financial and personnel support are necessary to implement these mandates.

Accountability, Oversight, Remedies

Given the large amounts of sensitive data being collected by governmental actors, accountability for breaches of trust are essential to incentivize the protection of privacy. Already, governmental actors are highly regulated, though much of this regulation is opaque to ordinary citizens and advocacy groups. One approach is to include patients and other stakeholders in the earliest phases of developing analytic programs, through community engagement boards or other consultative structures.[39] These structures should include interests aside from the government in order to be effective and represent a broad coalition of interested stakeholders, with funds provided to support community and non-profit involvement. A more active role for chief privacy officers and wider, more effective use of privacy impact assessments could also mitigate privacy risks.

The use of internal oversight mechanisms is of prime importance in order to protect patient data held by governmental entities. Without such oversight, ideally conducted by a chief privacy officer or similar staff member, there is an increased risk of inadvertent or intentional misuse of individual health data. The institutionalization of oversight procedures can also provide assurances to individuals that their privacy is being appropriately protected.

If violations occur, governmental entities should take steps to minimize any repercussions for individuals and their data records. These may include internal investigations, notifications to affected individuals, auditing, and external review.

Other governmental entities may also play a role in accountability and oversight mechanisms. Public legislative and administrative hearings, reports, and reviews by accountability boards can all provide oversight over government agencies that hold individual health data, and can serve as a valuable signaling function to the public that effective oversight programs are in place.

Conclusion

The enormous potential of health big data for governments is undeniable, and can be achieved in a privacy-protective way through responsible practices, such as those outlined in this paper. It is critical for governments seeking to use large amounts of citizen health information to institutionalize privacy-protective measures in advance of any collection and use of the data, guided by the FIPPs and by well-designed ethical rubrics. Government entities must carefully weigh the benefits and risks to using any kind of personal health information, provide contextual notice and transparency, and facilitate meaningful citizen engagement and government accountability for their data practices.

[39] See I. Glenn Cohen et al., *The Legal and Ethical Concerns That Arise from Using Complex Predictive Analytics in Health Care*, Health Affairs (July 2014) *available at* http://petrieflom.law.harvard.edu/resources/article/the-legal-and-ethical-concerns-that-arise-from-using-complex-predictive-ana.