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

Health Big Data in the Clinical Context

The health care industry, like other sectors, faces exciting new opportunities as a result of the cluster of developments occurring under the umbrella of the term "big data." Already, building on a tradition of research, health care providers, insurers, pharmaceutical companies, academics and also many non-traditional entrants are applying advanced analytics to large and disparate data sets to gain valuable insights on treatment, safety, public health, and efficiency. As they do so, they encounter privacy questions.

In some ways, the privacy issues surrounding research and other health uses of big data are not new. The limits of notice and consent have long been recognized. Issues of security plague even "small data." But health big data is more than just a new term. The health data environment today is vastly different than it was 10 years ago and will likely change more rapidly in the near future.

Most definitions of "big data" are based on the observation that the volume, velocity, and variety of data are rapidly increasing.[1] 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. Until recently, the health care sector had lagged in its use of information technology. However, that is rapidly changing, due to a variety of factors including the shift to electronic health records (driven in part by federal incentives) and the emergence of new ways to collect health data.[2] Hopes are high for big data as an important component of the "learning health care system," which aims to leverage health information to improve and disseminate knowledge about effective prevention and treatment strategies, to enhance the quality and efficiency of health care.[3]

---

[1] See Jane Sarasohn-Kahn, "Here's Looking at You: How Personal Health Information Is Being Tracked and Used" (CHCF, July 2014) http://www.chcf.org/publications/2014/07/heres-looking-personal-health-info. See generally, Executive Office of the President, "Big Data: Seizing Opportunities, Preserving Values" (May 2014) pp. 1-9.

[2] See Marc Berger, Kirsten Axelsen and Prasun Subedi, "The Era of Big Data and Its Implications for Big Pharma," Health Affairs Blog (July 10, 2014) http://healthaffairs.org/blog/2014/07/10/the-era-of-big-data-and-its-implications-for-big-pharma/. Participation in the electronic records incentives program under the HITECH Act of 2009 is running at over 90%, significantly exceeding initial predictions. See Centers for Medicare and Medicaid Services, Medicare & Medicaid HER Incentive Programs, HIT Policy Committee, September 3, 2014, http://www.healthit.gov/FACAS/sites/faca/files/HITPC_CMSDataReview_2014-09-03.pdf. (This figure is about providers who have registered for the program; it does not reflect those who have met the criteria and have begun receiving incentive payments.)

[3] See National Research Council, "*The Learning Health care System: Workshop Summary" (IOM Roundtable On Evidence-Based Medicine) (2007); Harlan M. Krumholz, "Big Data and New Knowledge in Medicine: The Thinking, Training, and Tools Needed For A Learning Health System," Health Affiars (July 2014)*. (The July 2014 issue of Heath Affiars is devoted to big data.)

To explore the privacy implications of health big data, and to develop concrete proposals for how to resolve privacy issues and at the same time reap the benefits of big data techniques, 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 first of these scenarios: clinical and administrative data generated by healthcare providers and payers in the course of providing treatment to patients, managing health care institutions, or processing payments. (This excludes data collected in controlled clinical trials, which we do not specifically address.[4]) We look both at big data uses by providers and payers and at their disclosures of data, when permitted, to third parties for research and other analytic purposes.

The Privacy Rule promulgated under the Health Insurance Portability and Accountability Act (HIPAA) covers most healthcare providers and payers. Some uses of clinical data for research are also covered under the "Common Rule," for protection of human research subjects in federally funded research.[5] There has long been concern about the limitations of both sets of rules and about inconsistencies between them.[6] Clearly, those two legal regimes should be more consistent; efforts to harmonize them have been launched but so far have not progressed.[7] The impetus of the big data revolution may spur harmonization and reform.

We acknowledge those long-running concerns, but in this paper we do not attempt to address them comprehensively or conclusively. Instead, we look to the framework provided by the Fair Information Practice Principles (FIPPs) and explore how it could be applied in an age of big data to clinical and administrative data.[8] The FIPPs informed, albeit imperfectly, the HIPAA Privacy Rule, just as they have influenced to varying degrees most modern data privacy regimes. While some have questioned the continued

---

[4] For recommendations on privacy and sharing data collected in controlled clinical trials, see "Sharing Clinical Trial Data – Maximizing Benefits, Minimizing Risks," Institute of Medicine Committee on Strategies for Responsible Sharing of Clinical Trial Data, National Academies Press (2015), http://www.iom.edu/Reports/2015/Sharing-Clinical-Trial-Data.aspx.

[5] See 45 CFR part 46, subpart A.

[6] Sharyl J. Nass et al., Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research (Institute of Medicine 2009).

[7] In July 2011, the U.S. Department of Health & Human Services (HHS) released an Advance Noticed of Proposed Rulemaking, seeking comment on potential changes to the Common Rule, some of which were designed to make HIPAA and that Rule more consistent. 76 Fed. Reg. 44512–44531 (July 26, 2011). However, HHS has published nothing further to advance that effort.

[8] The FIPPs are globally recognized as the foundation for information privacy. There is, however, no definitive version of the FIPPs.We use an articulation of the FIPPs drawn from three sources: the Markle Connecting for Health Common Framework (http://www.markle.org/health/connecting-health); the White House's 2012 Consumer Bill of Rights (http://www.whitehouse.gov/the-press-office/2012/02/23/we-can-t-wait-obama-administration-unveils-blueprint-privacy-bill-rights): and the ONC's Nationwide Privacy and Security Framework for Electronic Exchange of Individually Identifiable Health Information (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.

validity of the FIPPs in the current era of mass data collection and analysis, we consider here how the flexibility and rigor of the FIPPs provide an organizing framework for responsible data governance, promoting innovation, efficiency, and knowledge production while also protecting patient privacy. Rather than proposing an entirely new framework for big data, which could be years in the making at best, we believe that the best approach for data in the traditional health care system is to start with the FIPPs-based rules under HIPAA and the Common Rule and interpret them for big data uses. This effort could have the further benefit of laying the groundwork for a consistent set of principles covering both the traditional health sector and emerging consumer applications.[9]

Openness / Transparency

Despite the growing complexity of data flows within the health care field, the principle of openness or transparency is still relevant, both at the initial point of data collection and whenever patient data is used in subsequent analysis. Indeed, while many have said that the notice-and-consent model of privacy has failed, transparency about data practices is more important than ever.

As a fundamental principle, whenever data is collected about an individual, it should be clear to the individual what is being collected and how it will be used. Health care professionals who interact with patients are obligated to make sure that patients are informed about how their data will be used, including about the potential for secondary usage not directly related to the patient's treatment. The more unexpected or potentially objectionable a data collection or usage would be, the greater is the obligation to explain the practice to the patient. As Omer Tene and Jules Polonetsky have noted, "transparency with respect to the logic underlying organizations' data processing will deter unethical, sensitive data use and allay concerns about inaccurate inferences."[10]

Currently, the HIPAA Privacy Rule requires that covered entities provide patients with a notice of privacy practices (the NPP) that describes how their health information *may* be shared and used. However, HIPAA does not require notice of *actual* practices, so these notices are couched in hypothetical terms, providing little information about actual practices. To describe a breadth of functions—including those that are becoming more data intensive—these notices can use the term "health care operations," which few patients understand. HHS has sought to improve the quality of notices by developing model notices of privacy practices using plain language, approachable formats, and a

[9] The Federal Trade Commission, which has authority over otherwise unregulated consumer privacy practices, including health-related ones, also uses a FIPPs-based framework. See Federal Trade Commission, Protecting Consumer Privacy in an Era of Rapid Change: Reccomendations for Businesses and Policymakers, March 2012, http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.

[10] Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW J. TECH. & IP 239 (2013), http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1191&context=njtip.

layered approach.[11] However, the models still suffer from a lack of detail, using generic statements (for example, "We can use and share your health information to run our practice [and] improve your care . . . .") without encouraging covered entities to describe what sharing or analytic practices are actually in use. Indeed, the fact that all entities can comply with the HIPAA notice requirement by using the one-size-fits-all model is a disincentive to using the NPP to describe actual practices. And although the Office of the National Coordinator ran an innovative competition to develop online notices, it required the entrants to use the exact language of the paper-based model, thus failing to take advantage of the opportunities for more in-depth and dynamic disclosures online.[12]

The answer, however, is not to give up on transparency. Instead, the answer may be to distinguish between the HIPAA-required notice and transparency as to actual practices. Especially with the advent of big data, as the volume and uses of data grow, and as the HIT ecosystem becomes increasingly complex, providers and payers should reconsider what they disclose to patients, and how. Essentially, this means that, to build trust and improve accountability, regulated entities need to go beyond what HIPAA requires. One way to do this is to include in the paper-based notices concrete information about what entities actually do with patient data. Certainly, an entity should acknowledge in its NPP that it re-uses data for research purposes, and it could identify some major categories of research that it conducts. Another way to tell a fuller story of data flows is to supplement the notices provided to patients under HIPAA with an online document, which can be updated to provide a running list of actual research practices and other uses fitting within the broader categories of the paper notice. The mere process of drafting such a detailed description of actual practices can force an entity to inventory its data holdings and uses; once adopted, the statement can provide the foundation for internal and external accountability.

Providers may also consider whether there are lessons to be learned in terms of just in time delivery of information, which has been explored in the context of mobile apps.[13] HIPAA treats notice largely as a one-time event: the patient typically receives and signs a general document upon the first visit, which then remains valid for future collections and uses and is only required to be revised in the event of a material change in the rights and obligations described in the notice. Just-in-time disclosures could supplement the official notice of privacy practices with information in more digestible and context-specific increments.

When a patient sees a clinical provider, she likely expects that provider to collect sensitive health data. Each encounter offers an opportunity to provide digestible amounts of information regarding the collection and use of data. In the past, this has been difficult, given the time constraints on providers. The near ubiquity of smartphones

---

[11] United States Department of Health and Human Services, Model Notices of Privacy Practices, revised February 2014, http://www.hhs.gov/ocr/privacy/hipaa/modelnotices.html.
[12] Office of the National Coordinator, Digital Privacy Notice Challenge, http://oncchallenges.ideascale.com/a/pages/digital-privacy-notice-challenge.
[13] See Federal Trade Commission, Mobile Privacy Disclosures: Building Trust Through Transparency, February 2013, http://www.ftc.gov/sites/default/files/documents/reports/mobile-privacy-disclosures-building-trust-through-transparency-federal-trade-commission-staff-report/130201mobileprivacyreport.pdf.

may make it easy to deliver a link to an online disclosure providing more granular detail about data collection and use. Obviously, patients cannot be expected to understand the precise details of how every diagnostic machine works or the complexity of data flows in the health care learning system. Many patients will simply not be interested in detailed information about data practices. However, information about non-treatment uses should be spelled out in a notice that is somewhere accessible to patients, informing patients who are particularly interested and enhancing accountability.[14]

Payers and business associates that do not directly interface with patients should also clearly describe their data practices in transparency statements available on their websites. Certainly, few patients will directly access that information, but making it available to regulators, advocates, and interested patients will incentivize companies to adopt objectively reasonable business practices and will serve as a basis for holding them accountable to those policies.

True transparency may become more important because of another factor running in parallel with the development of the learning health care system: as more and more people get moved from employer-provided health insurance to purchasing individual plans – or in the case of employer-provided insurance, being more financially responsible for the costs of care -- they may end up with more of a typical consumer relationship with their health plans, and it will be in the interest of the health plans to become more transparent to strengthen that relationship.

Purpose Specification and Use Limitations / Respect for Context

Traditionally, the FIPPs were interpreted as requiring entities to specify up front the purposes for which data was being collected and to limit future use to those specified purposes unless new consent was obtained. However, the analytic capabilities of big data often are "aimed precisely at . . . unanticipated secondary uses."[15]

In the health care context, the purpose specification principle has proven especially problematic in limiting research uses of protected health information. Currently, under the HIPAA Privacy Rule, research is treated as a secondary use. Unless data are de-identified, under criteria spelled out in the Rule, or unless data are reduced to what is known as a "limited data set," individuals must give affirmative written consent

[14] While doctors have become extremely limited in the time they can spend with patients, they are not the only ones who interact with patients; there are nurses and often case managers (for people with chronic illness, for example). Plus, under new payment models, where providers are being paid for value and not per service, primary care physicians may need to spend more time with patients (especially those who have complex conditions to manage) and may need to explain to those patients the operation and role of monitoring devices. These and other changes in health care should be seen as opportunities to be more creative in accomplishing transparency on top of the HIPAA-required notice.

[15] Ira Rubinstein, *Big Data: The End of Privacy or a New Beginning?,* 3 Int'l Data Privacy Law 74 (2013), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2157659.

("authorization" in HIPAA terms) to research uses of their data.[16] Administratively, such authorizations are very hard to obtain, and essentially impossible after the data has been collected. Yet some of the most promising applications of health big data are in the field of research. There have long been concerns about the limits that the HIPAA Privacy Rule and the Common Rule impose on health research.[17]

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 research 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. This is an important protection that makes the limited data set very appealing.

HIPAA also allows research without authorization using de-identified data. Unfortunately, much controversy swirls around de-identification. Another part of the solution to the problem of research uses may be through improvements in the understanding of de-identification and the treatment of de-identified data (discussed below under "minimization"). The key is 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."[18]

For identified data, however, HIPAA-covered entities are tied to a notice-and-consent regime. We have recommended above that providers and payers should somewhere (probably online) make available specific information about actual uses of patient data for research purposes. We have noted that, if covered entities are not tied to the HIPAA NPP as the sole or even primary vehicle for transparency, one could expect more specific descriptions of research uses through other means. However, if we want to realize the benefits of big data, it will not be practical to make patients individually aware at the time of collection of the future, unanticipated uses of their data for research purposes and even after the fact it will not always be possible to tell a patient or plan member whether her information actually was used in particular research. Instead, other ways must be adopted to respect the principle of use limitation.

---

[16] There are some limited exceptions. The HIPAA Rule permits use or disclosure of personal health information for research if a waiver of authorization has been approved by an Institutional Review Board or privacy board meeting certain criteria. 45 CFR 164.512(i). It is essentially not possible to waive authorization for use of fully identifiable data in research; at least some identifiers need to be stripped out in order to qualify for an IRB waiver. However, waivers are granted in circumstances where the researcher needs more data than would be possible under the limited data set or to meet the de-identification standards, but some identifiers are still stripped out to reduce risk.

[17] For example, in a Hastings Center special report, *Ethical Oversight of Learning Health Care Systems*, renowned bioethicists challenged the traditional treatment of research uses of data and called for a new ethics framework for the learning health care system that would more expressly acknowledge contributing to learning from data as an ethical obligation of both providers and patients. *See* Nancy E. Kass et al., *The Research-Treatment Distinction: A Problematic Approach for Determining Which Activities Should Have Ethical Oversight*, 43 HASTINGS CENTER REP. 4 (2013); Ruth R. Faden et al., *An Ethics Framework for a Learning Health Care System: A Departure from Traditional Research Ethics and Clinical Ethics,* 43 HASTINGS CENTER REP. 16 (2013).

[18] Deven McGraw, Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data, June 26, 2012, http://jamia.bmj.com/content/early/2012/06/25/amiajnl-2012-000936.full.

Another part of the solution may lie in re-conceptualizing certain research uses. The 2012 White House report on consumer privacy uses "respect for context" as a way to describe the essence of the use limitation principle. Respect for context means that consumers have the right to expect that service providers will collect, use, and disclose personal data only in ways that are consistent with the context in which that data was provided. In the context of health big data, providers can use the initial context of data collection as a guide in circumscribing future uses of that data, while still allowing for innovative analytic practices – as long as those uses are related to the initial context.

While research (defined as "a systematic investigation . . . designed to develop . . . generalizable knowledge") is considered a secondary use under the HIPAA Privacy Rule, data usage for "operational" purposes is not treated as a secondary use and therefore requires no consent. But internal operations and research are often very similar in scope — with the only difference being whether the results (which are aggregate and not identifiable) are made publicly available. From a privacy perspective, this distinction is dubious: aggregate results (public or not) do not reveal the health conditions of identifiable patients.

Ideally, the rule sets for operational use and research would meet somewhere in the middle. As recommended above, health care providers and payers should be more specific about their operational uses in more detailed online notices, and, to the extent feasible, contextually as part of treatment encounters. And they should continue to be required to specifically disclose research uses, but those disclosures might be provided on their websites and by other means.[19] However, provided that more complete disclosure of actual research uses is available, health care facilities should not have to obtain specific authorization for research to evaluate the safety, quality and effectiveness of prevention and treatment activities conducted entirely within their organization (or by business associates acting on their behalf under strict data governance controls).

Under this approach, "context" would include future uses that could expand knowledge with regard to treatment of the disease the patient is suffering from, potentially advancing the patient's treatment as well as the treatment of others like her. In other words, internal usage by a provider (or a business associate under strict controls) for both service improvement and published research could be seen as logically consistent with the context in which the data is provided. This is consistent with recent changes to HIPAA's requirement for individual authorization for research. Historically, such authorizations needed to be study specific; but in 2013, guidance was issued allowing authorizations for future research to be more general, as long as the research was "sufficiently described" that a person would not be surprised to learn that her data were used for a particular research protocol.[20]

---

[19] One question is whether there is a threshold of scope or size for when research projects should be disclosed. At the least, projects that have been submitted to and approved by an Institutional Research Board should be disclosed. In the case of for-profit entities, proprietary issues may also have to be considered.

[20] In January 2013, HHS issued guidance on HIPAA's research provisions to make it easier to obtain patient authorization for research uses of data. 78 Fed. Reg. 5566-5702 (January 25, 2013).

One way to achieve this outcome is by a HIPAA amendment changing the definitions of research and operations to specify that "operational purposes" includes certain forms of research into treatment, efficiency and outcomes, even if the research is generalizable and published or that uses involving evaluation of the safety, quality and effectiveness of prevention and treatment activities do not constitute research requiring consent or IRB waiver.[21] Ethical concerns for potentially controversial studies could be mitigated by other oversight mechanisms.[22] It might be possible to achieve the same results though OCR guidance, similar to the guidance issued allowing more general consent for research uses of data. Under the system we propose, disclosure for research would still be subject to the minimum necessary standard and would be treated as "non-routine."[23]

On the other hand, the transfer of personal health information in identifiable form to third parties who are not business associates for non-treatment purposes does not respect context but rather would surprise most patients and should remain subject to limits. When a business associate agreement is in place, the current Rule already affords some flexibility for big data uses by permitting business associates to aggregate data from different covered entities, including data about the same patients in both sets, so long as covered entities do not receive back identifiable information about persons who are not their patients. That limitation remains valid.

Even when health data is not transferred to a third-party, providers and payers should be wary of leveraging patient data on behalf of external entities, especially for marketing purposes. Under amendments in the HITECH Act, providers cannot send marketing offers on behalf of third parties — targeted perhaps by information contained in a patient medical record — unless the patient affirmatively opts to receive these offers (unless they are communications about a drug the patient is already on (refill reminders, for example)).

Focused Collection / Collection Limitation

In the clinical health context, patients naturally assume that a tremendous amount of very personal information is going to be collected from them. On the other hand, health care providers should only collect the information reasonably necessary to achieve therapeutic goals. Absent specific consent to collect supplemental data for research, "treatment purposes" remains a valid limiting concept despite the increasing promise of

[21] In 2011, HHS issued an Advance Notice of Proposed Rulemaking that could have led to such a re-classification of research, but the process has not moved forward. In our view, it is urgent to restart the initiative.

[22] Lines may have to be drawn limiting research uses of more sensitive data (such as data commonly subject to enhanced protections) or prohibiting more unusual or potentially objectionable research practices.

[23] It may also be appropriate to address longstanding concerns with the potential over-breadth of the category of "operations." As big data becomes integrated into health care operations, it may be desirable to more clearly distinguish in the statute uses that improve quality and efficiency versus those that support marketing and fundraising.

secondary uses.[24] Medical facilities should not abuse the sensitive doctor-patient relationship to collect additional data that could hypothetically be relevant one day, either for the patient's care or for secondary usage. Data collection should be tied to the fundamental purpose of the visit or planned treatment regimen for the patient.

With the advent of big data, there has been some suggestion that limits on collection should be abandoned in favor of a focus only on preventing harmful uses of data. A moment's reflection, however, shows this not to be true. Collection of data in the clinical setting still must logically be subject to reasonable limitations. One fundamental rule that remains true even in the era of big data is that providers should not collect information about patients in ways that would surprise the patient.[25] In the health care context as in others, entities collecting data should consider the average person's reasonable expectations. An extreme example illustrates the point: patients in a hospital do not (and should not) expect that their phone calls with family members to be monitored — and it would be wrong for a hospital to record all such calls, even if the hospital did not misuse the data.

The era of big data poses special considerations as providers and payers increasingly may have the technical capacity to access external data sources about their patients. For example, doctors, other providers, and payers may increasingly seek access to health data generated by commercial gadgets and applications, such as Jawbone and Fitbit, or may enter into partnerships with such device and application developers. Care will need to be exercised in establishing these data flows to ensure that patients are aware of what data is being collected and how it is going to be used. Likewise, providers or payers could potentially access commercial data broker services (which providers and payers may already use for fraud control and to assess ability to pay and eligibility for certain forms of financial assistance) to glean information about patients' lifestyle choices and habits to inform care decisions. Other challenges arise with respect to collection of data that is publicly shared on social media or freely searchable on the web. There may be scenarios in the mental health field or other contexts where such information may be relevant to treatment. As a general rule, doctors should be extremely cautious about appending external data sources to clinical data, and should proactively seek out such information only with a patient's informed consent. Likewise, if payers become more interested in this data as they become more involved in managing chronic care, they should proceed cautiously, with full transparency and express consent. An IOM report[26] recommends the collection of socio-economic determinants of health by

[24] The 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 based on purpose.
[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." http://healthit.hhs.gov/portal/server.pt/gateway/PTARGS_0_0_6011_1815_17825_43/http%3B/wci-pubcontent/publish/onc/public_communities/_content/files/hitpc_transmittal_p_s_tt_9_1_10.pdf.
[26] The report is Capturing Social and Behavioral Domains and Measures in Electronic Health Records, Institute of Medicine Committee on Recommended Social and Behavioral Domains and Measures for Electronic Health Records, National Academies Press, 2014, http://www.iom.edu/Reports/2014/EHRdomains2.aspx.

healthcare providers, which we believe should only be done if patients are informed and consent to this type of collection.

Minimization

The principle of data minimization is relevant not only at the collection stage, but also when requesting or disclosing data to third parties. HIPAA's minimum necessary standard is quite explicit in limiting collection and disclosures, particularly for non-treatment purposes: "When using or disclosing protected health information [PHI], or when requesting PHI from another covered entity," a covered entity must make reasonable efforts to limit [PHI] to the minimum necessary to accomplish the intended purpose of the use, disclosure, or request."[27] In addition, under HIPAA providers cannot transfer clinical data about patients to third parties (non-business associates) for secondary uses such as marketing absent the patient's clear and informed consent ("authorization").

The advent of big data may prompt new ways of thinking about de-identification as a data minimization technique. Under HIPAA, data may be used and shared for secondary purposes if it has been meaningfully de-identified such that the information could not likely be traced back to a specific patient. It has been widely argued that de-identified data can be combined with other data to re-identify individuals, and this has led to criticism of reliance on de-identification as a privacy-protecting measure. We believe these concerns (which have mainly been raised with respect to data de-identified and used outside the HIPAA framework) point in the wrong direction. While there will always be a risk that de-identified data could be re-associated with particular patients, good faith, reasonable de-identification schemes, when coupled with enforceable limits against re-identification, represent a balanced approach to protecting personal privacy while still allowing commercial and scientific value to be extracted from data sets.[28]

The HIPAA Privacy Rule provides two methods for de-identifying health information – an expert determination method, by which an operator applies statistical or scientific principles to reduce the risk that an individual could be re-identified, and a safe harbor method, which mandates the removal of 18 types of identifiers and no actual knowledge that residual information could lead to re-identification. Questions have been raised about whether the safe harbor methodology for de-identification is sufficiently rigorous. At the same time, however, few entities use the statistical method, which may provide more protection while yielding greater data utility.[29]

The test promulgated by the Federal Trade Commission for data not regulated by HIPAA provides a similar, though more flexible, framework. That test states that data is not "reasonably linked" to an individual or a device if (1) the party takes reasonable

[27] 45 CFR 164.502(b).
[28] See Deven McGraw, "Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data," JAMIA (June 2012) http://jamia.bmj.com/content/early/2012/06/25/amiajnl-2012-000936.full.
[29] *See generally* Khaled El Emam, Guide to the De-Identification of Personal Health Information (2013); CDT BLOG, HIPAA de-identification (June 2009) https://www.cdt.org/files/healthprivacy/20090625_deidentify.pdf.

measures to de-identify the data, (2) commits to not re-identify data, and (3) prohibits downstream recipients from re-identifying the data.[30] If the HIPAA de-identification standards are found in practice to be too rigid, a shift to the FTC's objective standard may be appropriate.[31]

Research in effective anonymization should be supported. One promising line of research, already incorporated into the Google Chrome web browser, is RAPPOR (randomized aggregatable privacy-preserving ordinal response). RAPPOR is a technique that randomly changes the values of data while it is being collected, which has the attractive property of rendering the potentially-sensitive raw data meaningless -- it is randomized -- while allowing perfect precision in aggregated data, disconnected from the individual.[32] So far, to our knowledge, it has not been tested with health data.

De-identification cannot be presumed to eliminate all risk of re-identification of patient data; therefore, it is important to require assurances from recipients of de-identified data that the data will not be re-identified, and to provide penalties for re-identification.[33]

Another way to implement the minimization principle is to limit the centralization of data by linking distributed data networks in lieu of centralized collection of copies of data.[34] In the Mini Sentinel Distributed Database, which facilitates safety surveillance on drugs approved by the FDA, participating data sources put their data into a Common Data Model and perform the analytics; the aggregate results are then rolled up to produce the results (sometimes referred to as bringing the questions to the data).[35] Other models include allowing entities to make data available to others for analysis – such as by pushing data to a dedicated edge server- but without allowing the researcher to obtain a copy of the data. This enables analytics to be performed without releasing the raw data

---

[30] FTC, *"Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers,"* (March 2012) http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf "Reasonable measures," the FTC said, "means that the company must achieve a reasonable level of justified confidence that the data cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer, computer, or other device." The Commission went on to say that "what qualifies as a reasonable level of justified confidence depends upon the particular circumstances, including the available methods and technologies. In addition, the nature of the data at issue and the purposes for which it will be used are also relevant." "The standard," the Commission concluded, "is not an absolute one; rather, companies must take reasonable steps to ensure that data is de-identified. Depending on the circumstances, a variety of technical approaches to de-identification may be reasonable, such as deletion or modification of data fields, the addition of sufficient "noise" to data, statistical sampling, or the use of aggregate or synthetic data."

[31] Privacy consultant and consumer advocate Bob Gellman has proposed an alternative approach to de-identification proposal. Robert Gellman, *The Deidentification Dilemma: A Legislative and Contractual Proposal*;, 21 Fordham Intell. Prop. Media & Ent. L.J. 33 (2010), http://bobgellman.com/rg-docs/RG-Fordham-ID-10.pdf.

[32] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova, RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response (2014) http://arxiv.org/abs/1407.6981.

[33] Deven McGraw, *Building Public Trust in De-Identified Health Data*, 20 J. AM. MED. INFO. ASS'N 704 (2012), *available at* http://jamia.bmj.com/content/early/2012/06/25/amiajnl-2012-000936.full.html.

[34] Center for Democracy & Technology, Decentralizing the Analysis of Health Data, March 22, 2012, https://cdt.org/files/pdfs/Decentralizing-Analysis-Health-Data.pdf.

[35] Mini-Sentinel, About Mini-Sentinel, http://www.mini-sentinel.org/about_us/default.aspx (accessed June 29, 2013).

(a model that works particularly well for data sources without the expertise to perform the analytics). In the model most commonly employed by the Patient-Centered Outcomes Research Network, the analysis is "brought to the data" and only aggregate statistics returned.[36] Policies governing health big data should aim to provide incentives to use such privacy-enhancing technical architectures.

Generally, the data minimization principle also requires entities to delete data when it is no longer needed for a legitimate purpose. In the clinical context, data may need to be preserved for treatment purposes or to fulfill legal or ethical obligations. However, every data stewardship plan should include limits on data retention.[37]

Data Integrity and Quality / Data Access and Accuracy

Health care providers have an inherent incentive to ensure that patient health data is accurate. However, analytic techniques associated with big data make it possible in some contexts to draw useful insights or knowledge from data sets with many inaccuracies. On the other hand, big data analytics pose certain risks of confusing correlation with causation. Therefore, in the context of big data, the principle of data integrity or accuracy might be focused on the reliability or accuracy of outcomes.[38] All participants in the learning health system should be cognizant of the problem that big data often means more noise, not necessarily more signal. Selective publication of trial results by research institutions can amplify this problem considerably.

As HIPAA already requires, providers and payers should be obliged to provide patients with reasonable access to their health data.[39] There are a few reasonable exceptions to this requirement (for example, HIPAA excludes psychiatric notes from access requirements).

Developments occurring in tandem with the big data revolution may help realize the principle of access, which is a key part of the FIPPs framework. Access has been a right guaranteed in the HIPAA Privacy Rule, but never yet effectively implemented on a wide scale. That is rapidly changing. Beginning in 2014, providers participating in the meaningful use program are required to provide at least 50% of their patients with direct access to their clinical health information through portals to their provider's electronic

[36] PCORI, PCORnet: The National Patient-Centered Clinical Research Network, http://www.pcori.org/funding-opportunities/pcornet-national-patient-centered-clinical-research-network/.
[37] Many states have laws specifying minimum retention periods for health data, but generally these laws do not require deletion after the passage of any given time period. Nor does HIPAA specify limits on data retention.
[38] I. Glenn Cohen et al., "The Legal and Ethical Concerns That Arise from Using Complex Predictive Analytics in Health Care," Health Affairs (July 2014) ("rigorous validation of performance characteristics … will be particularly important").
[39] Business associate are not required to provide patients with a copy of their data unless the covered entity asks them to. Covered entities can delegate the access functionality, but they can't pass on the liability for failing to comply.

medical records.[40] Such access must include the capability to download this data into tools of the patient's choosing and to directly transmit the data to other entities.[41] In a related effort, HHS is promoting the Blue Button Initiative to expand 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 to empower consumers with easy and secure access to their health records from a variety of sources in formats they can use in conjunction with applications and services that help individuals to analyze data and make better use of it to manage their health. In September 2014, the Office of National Coordinator released a new toolkit to facilitate health industry adoption of the Blue Button functionality.[42] As providers and payers revamp their data flows to accommodate big data techniques, they might also embrace patient access. This might include giving patients access to the results of research undertaken with their data.[43] "If organizations provide individuals with access to their data in usable format, creative powers will be unleashed to provide users with applications and features building on their data for new innovative uses."[44]

Individual Participation / Control

With "research" currently defined broadly to include studies of treatment and effectiveness, both the HIPAA Privacy Rule and the Common Rule rely on patient consent to authorize and govern uses of data intended to expand the general knowledge base for effective and efficient care. Yet a disproportionate reliance on consent both limits very useful applications of data and may be less effective in protecting privacy.[45]

A thoughtful application of the FIPPs should include consideration of whether policies that enhance transparency to individuals about big data uses, that enable more active engagement and input of individuals in the research enterprise as well as access to research results, and that address security to protect against risks of internal misuse (such as unauthorized access) or inadvertent exposure of such data, would be more effective than mere consent at building public trust and protecting patient interests while facilitating health big data analytics. The relatively lessened role for individual control in the clinical health setting must be offset by a stronger application of the other Fair Information Practice Principles. These offsetting protections should include, *inter alia*,

---

[40] Centers for Medicare and Medicaid Services, Stage 2 Overview Tipsheet 4 (last updated August 2012), https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/Stage2Overview_Tipsheet.pdf, (accessed June 29, 2013).
[41] *Id.*
[42] http://bluebuttontoolkit.healthit.gov/selector/#intro-title.
[43] See Erika Check Hayden, "A Broken Contract," Nature (June 2012) (discussing the ethical questions in notifying study participants of results).
[44] Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW J. TECH. & IP 239 (2013), http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1191&context=njtip.
[45] For a more extensive discussion of the dangers of overreliance on consent, see CDT's January 2009 paper, "Rethinking the Role of Consent in Protecting Health Information Privacy," http://www.cdt.org/files/pdfs/20090126Consent.pdf.

more stringent restrictions on transfer of health information to third parties for secondary usage, more robust accountability regimes, and more rigorous security models.

In general, the HIPAA Rule already recognizes that, inside of the patient-to-provider context, individuals should not be expected to exercise meaningful control over their personal information — apart from the decision to allow it be collected in the first place. Consequently, a patient need not be notified every time that health information is transferred within an organization — or even outside an organization under the controls required in business associate agreements — for administrative, technical or other operational purposes. The guidelines under HIPAA appropriately allow for these types of uses without the active participation of the patient (while providing patients a limited right, after-the-fact, to obtain a list of some kinds of less common disclosures). Likewise, disclosure of de-identified data does not require consent.

On the other hand, the HIPAA Rule appropriately requires consent for the use and disclosure of identifiable patient data to third parties who are not business associates in most cases. In the age of big data, there are some areas arguably not covered by HIPAA where consent is necessary and effective. For example, if a provider or payer seeks to append information from non-clinical sources (e.g., from a data broker or personal health app), patients should have control over that information being collected and included as part of the medical record, and should be asked for their permission. Supplementing clinical data with outside sources — even public data — would surprise most patients, who typically do not expect that their doctors (or payers or service providers) will be collecting information about their behavior outside the clinical context.

Security

Under the HIPAA Security Rule, healthcare providers and payers must have in place policies and procedures, and must appropriately train personnel, to protect the security of electronic personal health information. Researchers and others receiving data under business associate agreements are likewise required to adopt security protections. All measures must be commensurate with and calibrated to privacy risks and, specifically, scaled to identifiability. As a result of the HITECH Act, HIPAA includes a breach notification rule, requiring providers to alert patients when their data has been lost or stolen.

However, existing practices are likely to be insufficient as the healthcare system, especially with its growing emphasis on learning and big data, involves increasingly complex data flows. As cybersecurity guru Bruce Schneier has often noted, the more complex an information system is, the harder it is to secure.[46] If "the breach at **Target Corp.** that exposed credit card and personal data on more than 110 million consumers appears to have begun with a malware-laced email phishing attack sent to employees at an HVAC firm that did business with the nationwide retailer,"[47] then healthcare providers

[46] Kim Zetter, *Three Minutes With Security Expert Bruce Schneier*, PC WORLD, September 28, 2001, https://www.schneier.com/news-038.html.
[47] Brian Krebs, *Email Attack on Vendor Set Up Breach at Target*, KREBS ON SECURITY, February 14, 2014, http://krebsonsecurity.com/2014/02/email-attack-on-vendor-set-up-breach-at-target/

and payers, with complex relationships with dozens or hundreds of vendors, partners and affiliates, face truly daunting risks. The sharing of data to take advantage of big data applications only further exacerbates those risks.

Unfortunately, the healthcare sector appears to be lagging on security.[48] This is probably not due to a lack of guidance as to security standards. The Office of Civil Rights at HHS has provided guidance[49] and so has the National Institute of Standards and Technology.[50] Instead, it appears that some, possibly many covered entities are not devoting sufficient attention or resources to security. Standards such as the ISO/IEC 27799 information security guidelines in health are not widely implemented. There have been many reports of breaches of unencrypted data, even though encryption of data at rest and in transit is an "addressable implementation specification" under the Security Rule, meaning that HIPAA-covered entities are expected to implement it unless it is not "reasonable and appropriate" to do so. The HHS Office of Civil Rights, which has enforcement responsibility, last year conducted its first-ever audits and found that a lot of entities just do not pay sufficient attention to compliance with the Security Rule.[51] In the HITECH Act, State Attorneys General were granted authority to enforce HIPAA, but few of them have taken up the mantle (probably because they have limited resources and other consumer protection priorities).

Accountability, Oversight, Remedies

Either the industry must substantially improve its security practices, or HHS (and other regulators) will have to become more regulatory (something it is not well-suited for when it comes to information security) or more aggressive in seeking sanctions. Ideally, the government and industry should work together to improve practices. Policymakers promoting the benefits of big data should emphasize the need to address security in designing sharing and analytics programs. The de-identification, anonymization, and decentralization techniques discussed above should be part of such design decisions.

One of the most important requirements imposed by the HIPAA Security Rule is that covered entities must conduct an assessment of potential risks and vulnerabilities to the confidentiality, integrity and availability of the electronic health data they hold, to provide a basis for adoption of appropriate safeguards. Even in the absence of specific security standards, the risk assessment can be an important tool. Risk assessment should be

---

[48] *See*, *e.g.*, Ponemon Institute, Fourth Annual Benchmark Study on Patient Privacy and Data Security, March 2014, https://www.privacyrights.org/sites/privacyrights.org/files/ID%20Experts%204th%20Annual%20Patient%20Privacy%20&%20Data%20Security%20Report%20FINAL.pdf.

[49] United States Department of Health and Human Services, Health Information Privacy: Security Rule Guidance Material, http://www.hhs.gov/ocr/privacy/hipaa/administrative/securityrule/securityruleguidance.html.

[50] NIST Special Publication 800-66, "An Introductory Resource Guide for Implementing the HIPAA Security Rule," http://www.hhs.gov/ocr/privacy/hipaa/administrative/securityrule/nist80066.pdf.

[51] Brian Fischer, *HIE Security, Risk Analysis in the Spotlight*, SECURITY COMPLIANCE ASSOCIATES, August 8, 2013, http://www.scasecurity.com/hie-security-risk-analysis-in-spotlight/ (detailing Health IT Policy Committee recommendations made because of failures revealed by initial OCR security audits).

ongoing, but an especially good time to conduct or refresh a risk assessment is when providers and payers are planning the introduction of big data analytic techniques, with attendant changes in data flows.

As the size and diversity of datasets grow, and as analytic techniques make it easier to re-identify data and draw inferences from seemingly innocuous data, internal and external accountability must play a larger role in protecting clinical health data. Regulators should expand enforcement efforts aimed at obtaining sanctions whenever data is illegitimately used or transferred, or reasonable security protections are not implemented.

New governance structures should be explored and adopted. 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.[52] A more active role for chief privacy officers and wider, more effective use of privacy impact assessments could also mitigate privacy risks.

Conclusion

Building and maintaining public trust in a broader, robust, health big data ecosystem will require the development and implementation of comprehensive, adaptable policy and technology frameworks. Such frameworks should:

- provide protections for health data while still enabling analytics to solve pressing health challenges;
- apply consistently to health data regardless of the type of entity collecting it (be it a hospital or a commercial health app) and yet still be flexible enough to respond to the particular risks to privacy posed by different health data sharing models;
- include mechanisms to hold entities collecting and analyzing health data accountable for complying with rules and best practices;
- provide incentives for the adoption of privacy-enhancing technical architectures/models for collecting and sharing data; and
- be based on thoughtful application of the Fair Information Practice Principles (FIPPs), which have been the foundation for privacy laws and industry best practices both in the U.S. and internationally.

While the possibility of big data analytics holds great promise, clinical providers should take care to ensure that such uses take into consideration the sensitivity of patient data. Entities should follow the full complement of fair information practices in using identifiable data for these purposes, including (but not limited to) being transparent with patients about how their data is used for treatment and quality, safety and effectiveness evaluation purposes; using only the minimum amount of data needed to accomplish the

[52] 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).

particular activity; and protecting the data with security measures that are commensurate with the risks to privacy.