nature medicine

FOCUS | REVIEW ARTICLE

https://doi.org/10.1038/s41591-018-0272-7

Privacy in the age of medical big data

W. Nicholson Price II[1,2,3] and I. Glenn Cohen[2,3,4]*

Big data has become the ubiquitous watch word of medical innovation. The rapid development of machine-learning techniques and artificial intelligence in particular has promised to revolutionize medical practice from the allocation of resources to the diagnosis of complex diseases. But with big data comes big risks and challenges, among them significant questions about patient privacy. Here, we outline the legal and ethical challenges big data brings to patient privacy. We discuss, among other topics, how best to conceive of health privacy; the importance of equity, consent, and patient governance in data collection; discrimination in data uses; and how to handle data breaches. We close by sketching possible ways forward for the regulatory system.

Big data has come to medicine. Its advocates promise increased accountability, quality, efficiency, and innovation. Most recently, the rapid development of machine-learning techniques and artificial intelligence (AI) has promised to bring forth even more useful applications from big data, from resource allocation to complex disease diagnosis[1]. But with big data comes big risks and challenges, among them significant questions about patient privacy. In this article, we examine the host of ethical concerns and legal responses raised. Nevertheless, attempts to reduce privacy risks also bring their own costs that must be considered, both for current patients and for the system as a whole.

We begin by discussing the benefits big data may bring to health science and practice and then turn to the concerns big data raises in these contexts. We focus on a prominent (but not the only) worry: privacy violations. We present a basic theory of health privacy and examine how privacy concerns play out in two phases of the life cycle of big data's application to health care: data collection and data use. We ground these concerns in a discussion of relevant US law, a key feature of the health data world faced by innovators in this space, and make some regulatory recommendations. We argue, counter to the current zeitgeist, that while too little privacy raises concerns, it is also true that too much privacy in this area can pose problems.

Why do we need big data in health?

Big data has long been promised to substantially improve health care. But what is big data and why does it matter? Big data is often defined by 'three Vs': volume (large amounts of data), velocity (high speed of access and analysis), and variety (substantial data heterogeneity across individuals and data types), all of which appear in medical data[2]. We can organize the research applications of big data into two rough groups: long-practiced analysis approaches and newer methods using machine learning and AI.

Big data enables more powerful evaluations of health care quality and efficiency, which then can be used to promote care improvement[3]. Currently, much care remains relatively untracked and underanalyzed; amid persistent evidence of ineffective treatment, substantial waste, and medical error[4], understanding what works and what doesn't is crucial to systemic improvement. Big data can help: it can be leveraged to measure hospital quality, as in the Centers for Medicare and Medicaid Services' Hospital Inpatient Quality Reporting program[5]; to develop scientific hypotheses, as with proliferating genome-wide association studies[6]; to compare the effectiveness of different interventions, as in the Patient Centered Outcome Research Institute (http://www.pcori.org); and to monitor drug and device safety, as with the Food and Drug Administration (FDA) Sentinel system[7].

A new set of tools that use AI techniques to find patterns in big health data, which then can be used to make predictions and recommendations in care, is rapidly developing[8,9]. The best-known of these tools involves image analysis and is beginning to enter clinical use. Algorithms have been able to identify cancerous skin lesions from images as accurately as trained dermatologists[10], and the IDx-DR system has recently received FDA approval for image-based AI diagnosis of diabetic retinopathy. Further afield, AI can be used for prognostic purposes—to predict when trauma patients are about to suffer a catastrophic hemorrhage and need immediate intervention[11] or when patients are very likely to die within a year and therefore might consider shifting from traditional care to palliative care[12].

AI algorithms could also make recommendations for treatment (Box 1). Finally, and somewhat controversially, AI algorithms could help make resource-allocation decisions (Box 1)[1]. All of these uses require very large sets of health-care data, including how patients have been treated, how patients have responded to treatment, and personal patient information, such as genetic data, family history, health behavior, and vital signs[13]. Without these data, algorithms cannot be trained or evaluated on how they perform following training[14].

The next evolution of big data in health care—which is slowly gaining momentum—lies in the development of learning health systems[15]. In learning health systems, the traditional boundary between clinical research and care is eroded—although even in more traditional health system designs, there is significant fuzziness, doubt, and gamesmanship regarding the line between 'research' and 'quality improvement' or 'innovation', with important ramifications for regulatory review[16–20]. In learning health systems, data are collected routinely in the process of care, with the explicit aim that those data be used for the purpose of analyzing and improving care. Just as data are continuously collected, they are continuously analyzed to reveal patterns in the process of care, procedures that can be improved, and other underlying patterns such as differential patient response to different treatments[21]. Finally, these new insights are routinely incorporated back into the clinical care pathway, whether explicitly (in practice guidelines or publications) or implicitly (in the context of recommendations or procedures automatically embedded into

[1]University of Michigan Law School, Ann Arbor, MI, USA. [2]Project on Personalized Medicine, Artificial Intelligence, & Law, Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics, Cambridge, MA, USA. [3]Center for Advanced Studies in Biomedical Innovation Law, University of Copenhagen, Copenhagen, Denmark. [4]Harvard Law School, Cambridge, MA, USA. *e-mail: igcohen@law.harvard.edu

Box 1 | Vignettes illustrating possible uses of big data

- Scott suffers from liver cancer. Anita, his physician, is deciding which chemotherapeutics to administer. She turns to the CancerChoice module in the hospital's EHR system. This module pulls data from Scott's EHR—his medical history, family history, and genetic sequence—but also automatically links to large collections of commercially collected data to acquire additional information about Scott's shopping, eating, and exercise habits, which can help inform treatment choice. The module then makes a recommendation by combining all the data it has gleaned about Scott with similar data—both health-care data and health-related lifestyle data—from millions of patients across the country.
- Samantha presents at Chicago Hope Hospital with moderate organ dysfunction. The physician is trying to decide whether to send Samantha to a specialized intensicve care unit (ICU); Samantha might benefit, but beds are limited and other patients might benefit more. In traditional medicine, assessing a patient's risk for cardiopulmonary arrest or other preventable serious adverse events might take hours; furthermore, the assessment also has limited prognostic accuracy, and the risk may change during that period. Imagine that an alternative assessment mechanism is available. CorazonAI has developed a predictive analytics engine, based on data from millions of US patients' EHRs, that could ascertain the risk accurately for hundreds of patients with real-time updates to help the physician evaluate who should be admitted to the ICU[1]. The physician uses the system, which advises that Samantha be admitted.

In these vignettes, have patients' privacy been violated? Are these violations unethical? Are they ones the law should prohibit? And how do these concerns stack up against the benefits obtained from using big data in the health context?

electronic health record (EHR) systems). The concept of a learning health system can be applied either through explicit learning mechanisms or through AI algorithms, though at least for the foreseeable future we would expect humans to remain embedded firmly within the loop of learning–analysis–implementation.

How to think about health privacy

The concept of privacy is notoriously difficult to define. One currently prominent view connects privacy to context. There are contextual rules about how information can flow that depend on the actors involved, the process by which information is accessed, the frequency of the access, and the purpose of that access[22,23]. When these contextual rules are contravened, there has been a privacy violation. Such violations can occur because the wrong actor gets access to the information, the process by which information may be accessed is violated, or the purpose of access is inappropriate, and so on. Normative reasons why such violations are problematic can be divided (with some simplification) into two categories—consequentialist and deontological concerns. Two caveats are in order: first, some privacy violations raise issues in both categories. Second, some concerns we discuss are also present for 'small data' collection. Big data settings, however, have a tendency to increase the number of persons affected, the severity of the effects, and the difficulty for aggrieved individuals to engage in preventive or self-help measures.

Consequentialist concerns. Consequentialist concerns result from negative consequences that affect the person whose privacy has been violated. These can be tangible negative consequences—for example, one's long-term-care insurance premium goes up as a result of additional information now available as a result of a breach of privacy, one experiences employment discrimination, or one's HIV status becomes known to those in one's social circle—or these can be the emotional distress associated with knowing that private medical information is 'out there' and potentially exploited by others: consider the potential for increased anxiety if one believed one was now susceptible to identity theft, even before any misuses of identity have occurred (Fig. 1).

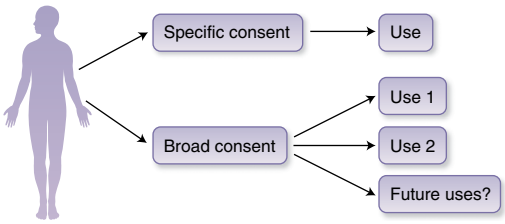

Fig. 1 | Consent models for health data. Specific consent allows individuals to control each specific use of their data. In broad consent, individual give blanket consent for all uses of their data.

Deontological concerns. Deontological concerns do not depend on experiencing negative consequences. In this category, the concern from a privacy violation manifests even if no one uses a person's information against this person or if the person never even becomes aware that a breach has occurred. One may be wronged by a privacy breach even if one has not been harmed. For example, suppose that an organization unscrupulously or inadvertently gains access to data you store on your smart phone as part of a larger data dragnet. After reviewing it, including photos you have taken of an embarrassing personal ailment, the organization realizes your data is valueless to them and destroys the record. You never find out this happened. Those reviewing your data live abroad and will never encounter you or anyone who knows you. It is hard to say that you have been harmed in a consequentialist sense, but many think the loss of control over your data, the invasion, is itself ethically problematic even when harm is absent. This is a deontological concern (Fig. 1).

Gathering data

Custodian-specific versus blanket provisions. The gathering of medical data raises many legal and ethical privacy questions. We focus here on the treatment of health data in the United States, but it is worth comparing the US approach with the EU approach[24]. Health data come from many different sources: EHRs, insurance claims, Internet of things devices, and social media posts, to name but a few. US privacy law treats health data differently depending on how they are created and who is handling the data—that is, who is the custodian. By contrast, the EU General Data Protection Regulation sets out a single broadly defined regime for health data (as well as other data), no matter what format, how it is collected, or who the custodian is[25]. It defines the category of 'data concerning health' broadly to mean "personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status."[26]

The custodians that US law focuses on are physicians, health systems, and their business associates. The major US federal law governing health data privacy is the Privacy Rule created under the Health Insurance Portability and Accountability Act (HIPAA)—there are also state-specific privacy laws and the federal Common Rule, which protects research subjects, but they are not our focus here[27].

Under the HIPAA Privacy Rule, 'covered entities' are prohibited from using or disclosing 'protected health information' (PHI) except in a specified list of circumstances; 'business associates' face similar limitations under required contracts with covered entities[28]. The definition of PHI is broad, including most individually identifiable health information; covered entities includes most health care providers, health insurance companies, and 'health information clearinghouses.'[28]

HIPAA creates a set of rules that are arguably both overprotective and underprotective of privacy (HIPAA also directly protects information security through a separate Security Rule[29]). On the overprotective side, while HIPAA allows use of PHI for health care treatment (including 'quality improvement'), operations, payment, public health, and law enforcement—it does not allow the use of PHI without Institutional Review Board (IRB) waiver or patient authorization for research, which is to say the systematic production of generalizable knowledge (Fig. 2)[30].

As to health data covered by HIPAA, the rule also has gaps. One of HIPAA's most important strategies for protecting patients from privacy breaches while enabling data sharing is deidentifying their data by removing a set of 18 specified identifiers, like names and email addresses[31]. However, deidentified data may become reidentifiable through data triangulation from other datasets (Box 2)[24,32–34]. Moreover, HIPAA focuses its regulation on particular actors and their activities, not the data themselves. For instance, once a patient requests their own health data—which HIPAA gives them the right to do, and some concerted efforts encourage patients to do[35,36]—if the patient then gives those data to another individual, HIPAA does not restrict use or disclosure of those data (unless the recipient is another covered entity or a business associate)[24].

But the more fundamental problem is that the majority of health data is not covered by HIPAA at all (Fig. 2). When Congress enacted HIPAA in 1996, it envisioned a regime in which most health data would be held in health records, and so it accordingly focused on health care providers and other covered entities. In the big-data world, the type of data sources covered by HIPAA are but a small part of a larger health data ecosystem. HIPAA does not cover health care data generated outside of covered entities and business associates, such as health care–related information recorded by life insurance companies. It does also does not cover health (as opposed to health care) data generated by a myriad of people or products other than the patient. It does not cover user-generated information about health, such as the use of a blood-sugar-tracking smartphone app or a set of Google searches about particular symptoms, and insurance coverage for serious disorders. And it certainly does not cover the huge volume of data that is not about health at all, but permits inferences about health—such as the information about a shopper's Target purchases that famously revealed her pregnancy[34,37,38]. This focus on data specifically arising from health 'care' contrasts with European regulation of data concerning health more generally.

We are already entering a future in which traditional health care spaces, HIPAA's covered entities, are being supplanted in the health data space by behemoths like Google, Apple, and IBM—all of which operate outside of HIPAA's regime. While, as we discuss below, some laws may protect particular uses of those data, overall there is little to protect patients from these threats to their health privacy in the United States at the moment.

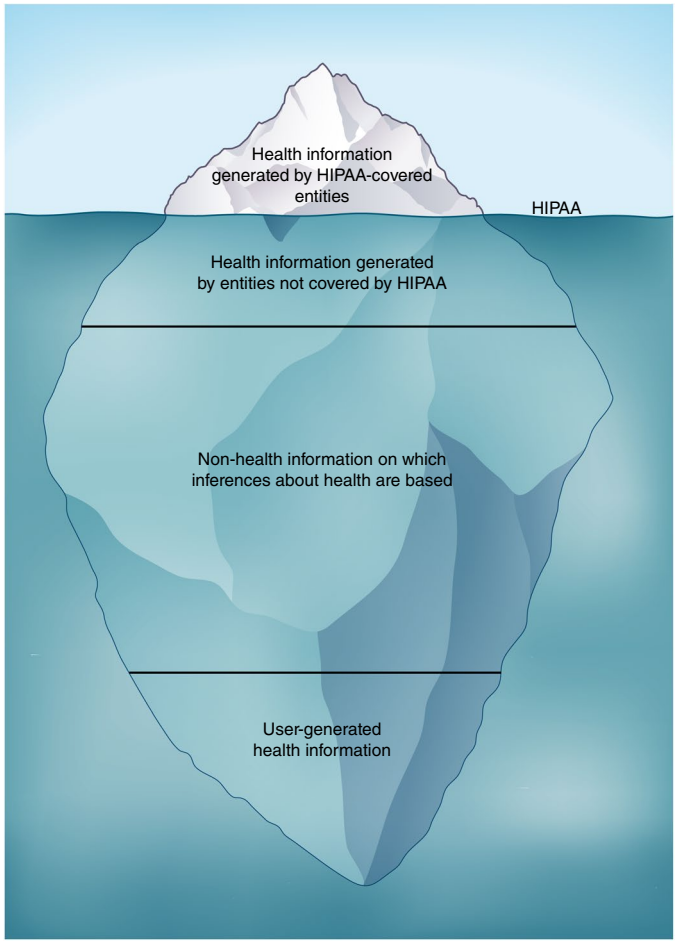

Fig. 2 | The data that is and isn't included in HIPAA. Information above the water indicates that information covered by HIPAA, and information below the water is not covered.

Equitable data collection. Another concern is not that too much data is taken from patients, but that data collection is not occurring equitably. As an ethical matter, data collection is best justified as a kind of 'bargain' struck between data sources and data users—sources provide users with data, recognizing that this may encroach in some ways on source privacy, because it will permit the users to provide advances in health care that will improve the lives of sources. When this balance is off, the bargain may break down. Existing bias can reappear in data mining, as has been shown for predictive analytics in policing: racial disparities in policing patterns result in racially biased predictions of criminal activity[39]. Unfortunately, health data have many of the same problems. Marginalized populations that are missing from non-health data such as credit card use or Internet history—leading to biases in credit scores or consumer profiles—may also be absent from big health data, such as genomic databases or EHRs, due in part to lack of health insurance and the inability to access health care as well as a number of other reasons[40]. The distributional consequences of this lack of inclusion in big data are complex; in some instances it may favor but in other instances disfavor those whose data are missing. For example, consider an allocation decision between multiple patients as to a scarce medical resource. If a particular minority group actually responds less well to the medical intervention then other groups, failure to collect information on the minority group might lead the algorithm to give the minority patient more priority than had the data been included. If the minority group responds better than other groups, the opposite effect might result. Whichever way it cuts, though, the result will be that the system's prediction will be problematically biased. This is a hard nut to crack, among other reasons because of contested and incompatible definitions of 'fairness' in the predictive analytics space[41,42]. One solution, is better access to health care for underserved populations, but if that goal is reached it will not be because of big data needs. Statistical adjustment for data gaps may help mitigate the problem somewhat, but

this is an area to which funders (especially public funders) should be attuned. The All of Us program, for instance, aims to develop a nationally representative sample for its genomic work (https://allofus.nih.gov/about/about-all-us-research-program). While that ambition will not be realizable for all big data research, funders should consider asking applicants to explicitly address their strategies for making their datasets more inclusive, and they should take that into consideration in allocating grants.

The role of the patient in data collection and access. To what extent should an individual's data be available for use in predictive analytics without her consent (Fig. 3)? An example is the use of EHRs without consent to build the proprietary CancerChoice model discussed in Box 1. Especially for deontological concerns with health privacy, the loss of control over who accesses an individual's data and for what purpose matters, even if there are no material consequences for the individual or if the individual does not even know.

Should some health data be seen as a kind of public good that can be conscripted for some potentially publicly minded uses? Here the notion of privacy as stemmed from contextual rules, discussed above, is particularly helpful. The ethical analysis will depend heavily on the type of data, including its identifiability; who will be accessing it; and for what purpose. Take one data source, EHR data stripped of the 18 HIPAA identifiers. One might feel differently about the Center for Disease Control (CDC) accessing this data for flu-tracking purposes compared to a hospital system using it to reevaluate its staffing and workflow to improve both cost efficiency and patient experience or to a pharmaceutical company using it for product development. Even if privacy is violated, it may be that, all things considered, the violation is outweighed by equitably distributed benefits in some cases. As a guiding principle for this analysis, one might think that individually unconsented use is more appropriate (especially for relatively deidentified data) the more the contributing patient will benefit from the data use—a principle of reciprocity—and where the risks to the patient (including the consequentialist risks discussed below) are low, such that the 'ask' of patients is small compared to the benefit—a principle of proportionality[16,43].

Second, whether or not patients consent for their data to be included within a set, what role should they have in deciding what kind of uses of their data are permissible? This is a question of designing a governance regime—and it matters to patient privacy because, as discussed below, many of the privacy harms of big health data arise not merely in the collection of data, but in their eventual use. On one extreme, one could imagine enabling every patient to approve every access to every piece of data individually after a purpose has been stated—a regime that would maximize patient autonomy but could eliminate most work using big health data[44].

On the other extreme, one could treat data as completely 'alienable,' such that the patient retains no rights of control, whether by external mandate or by 'broad consent,' as has been proposed in the biobank context[45,46]. As noted, our conception of privacy is contextual and the analysis will depend on the specifics of who seeks access to what data in what way for what purpose. For many cases, though, the optimal governance regime may lie somewhere in the middle. This might involve, for example, chartering a steering board that includes representative patients in deciding which requests for data to permit and under what circumstances. One analogy would be the Independent Review Panels that have been used to approve or deny requests for the sharing of clinical trial data[47]. A slightly different approach would be to actually put the data in a charitable trust, with trustees (some of whom would be patient representatives) making decisions about access conditions and approved uses while owing fiduciary duties to the patients' whose data is used, a model championed by some for biobanks[48].

Box 2 | The challenge of multiple data sets for reidentifiability

Many assume that 'anonymized' data cannot be used to reidentify the subject of the data. Unfortunately, as data sets proliferate, the ability to combine multiple datasets may defeat the deidentification strategy. The most famous example, which preceded HIPAA, was demonstrated by Latanya Sweeney. In the 1990s, the state of Massachusetts purchased health insurance for state employees and subsequently released records summarizing every state employee's hospital visits at no cost to any researcher who requested the data. Then-Governor William Weld assured the public that the data had been scrubbed to defeat reidentification by removing information such as names, addresses, and Social Security numbers. Unfortunately, many patient attributes were not scrubbed. Sweeney, then a graduate student, knew Weld resided in the city of Cambridge, and so she purchased this city's complete voter rolls, which contained the name, address, ZIP code, birth date, and sex of every voter in the city. She paired that data with the state health insurance data to demonstrate that one could reidentify Weld's prescriptions, diagnosis, and medical history[72,73].

A more recent example of the same problem outside of medicine pertains to the prize offered by Netflix in the mid-2000s to improve its movie recommendation algorithm. To enable the competition, Netflix publicly released one-hundred million records revealing hundreds of thousands of user ratings from 1999 to 2005. Netflix stripped identifying information but added unique user numbers to group ratings by users. Two researchers from the University of Texas, Arvind Narayanan and Vitaly Shmatikov, showed that one could nonetheless reidentify Netflix users by linking to other datasets. In particular, they drew on the publicly available data from the Internet Movie Database (IMDb), wherein users also rate movies but do so publicly, to offer a proof of concept. They showed that "Given a user's public IMDb ratings, which the user posted voluntarily to reveals some of his … movie likes and dislikes, we discover all ratings that he entered privately into the Netflix system." In particular, their reidentification strategy took advantage of ratings for more obscure movies in both systems and also the timing in which reviews were posted.

To be sure, neither of these examples are meant to show that deidentification is never possible or that reidentification will always be easy. Instead, they are meant to show how the increase in the number of datasets and linking of information makes reidentification more plausible even for data that had otherwise been thought deidentified[73,74].

Still another approach is what Barbara Evans, a law professor at the University of Houston, calls 'consumer-driven data commons,' "institutions that enable groups of consenting individuals to collaborate to assemble powerful, large-scale health data resources for use in scientific research, on terms the group members themselves would set."[49] There are many other governance possibilities[50], including so-called 'citizen juries' that have been used in the United Kingdom in these domains[51]; but especially where individualized patient consent will not be collected, it is important to have patient representatives involved in crucial decisions about how their data will be used.

While approaches built on any of these models may be feasible at the current moment, they may be less feasible in a future where datasets—containing not only huge amounts but huge varieties of data—are used for multiple different analyses. Such cross-context datasets and data-uses—using collections of consumer data to make health predictions, collections of health data to target advertising, or

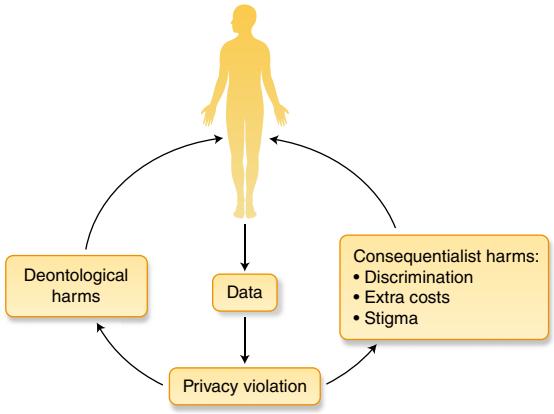

Fig. 3 | Potential harms to the individual if data is breached. The types of harm that can befall an individual once their data is leaked.

joint collections for both purposes—would make it harder to meaningfully set one governance regime for consumer data and another for health data. And to the extent that policymakers today require context-specific regimes, they may limit exactly that future development of cross-context datasets, for good and ill.

Data uses

In this section, we outline major legal and ethical privacy issues raised by using already-collected patient data, especially in AI-driven systems, and approaches for addressing them.

Discrimination based upon health data. The use of patient-derived big data in medicine can lead to consequentialist privacy concerns. One well-characterized set of objective harms comes from the possibility of discrimination: if employers or insurers learn of sensitive patient information from medical data, such as a debilitating or expensive disease, they may wish not to employ or insure that person, especially since in the United States health insurance is typically tied to employment[52]. Some would argue that this type of discrimination is justified under a principle of 'actuarial fairness,' where everyone should pay or be paid according to their risk as precisely as possible[53]—an enterprise that big data could make much easier. This raises a very fundamental question about whether to favor a notion of 'to each according to his risk' as opposed to a more solidaristic view of insurance, whereby to some extent we redistribute through insurance pooling[54]. In any event, our existing laws in health insurance and employment contexts have favored the latter view, prohibiting some but not all of this sort of discrimination.[55]

The Genetic Information Nondiscrimination Act (GINA) prohibits discrimination by health insurers or employers on the basis of genetic information, the Americans with Disabilities Act (ADA) prohibits discrimination in employment and insurance based on medical conditions that are disabilities, and the Patient Protection and Affordable Care Act (PPACA) prohibits health insurance discrimination and sharply limits medical underwriting. These laws represent an attempt to limit consequentialist privacy harms by limiting consequences of access to data rather than focusing on protecting data themselves (though GINA does also include some limits on data acquisition).

But these laws have important limits. The ADA, for example, will not limit uses of big data to adversely treat "people who are currently healthy but are perceived as being at high risk of becoming sick in the future."[56] Neither GINA nor the ADA reaches life insurance. And even when these laws do apply, they can be hard to enforce because it is often hard to know when discrimination has occurred. Moreover, other kinds of consequentialist harms are hard to address through law at all, such as stigma that can arise from others knowing about a sexually transmitted infection or learning that a child's parent is not the child's biological parent.

A recent survey of clinical trial participants on the sharing of participant-level clinical trial data beyond genomic information found that 6.6% were "very concerned" and 14.9% were "somewhat concerned" that "I could be discriminated against if the information was linked back to me," but as the authors acknowledge, specific characteristics of that study population, especially the fact that they have already decided to participate in clinical trials, may make it a poor predictor for general public attitudes on these questions[57,58].

Sharing of private information. A second set of consequentialist privacy harms involves more subjective injuries. Patients whose private health information becomes available can suffer embarrassment, paranoia, or mental pain. Even though these injuries may not have measurable external effects—the patients may suffer no financial injury or encounter no stigma from others—they are still injuries[59]. Laws like GINA, the ADA, or the PPACA have little purchase on this type of injury.

Big data also raises the possibility of more dignitary harms. In order to live a flourishing life, it is important that there be a part of an individual's life that is his or hers alone, that remains unknown to others unless shared. Facts about health are particularly sensitive and private. In some instances, big data permits direct knowledge regarding a person's health by others whom the individual would not want to access the information—whether through inadvertent disclosure or malicious activities such as hacking. Most people are woefully unaware of the uses to which their data may be put; a particularly salient example comes from use of the GEDmatch genetic database to help identify the Golden State Killer[60]. This example also helpfully illustrates the problem that information shared about one individual may reveal information about other individuals—here, genetic relatives—who are unaware that potentially revealing information has been shared and who have not consented to the sharing.

A more subtle and more difficult issue raised by predictive analytics is whether a person's privacy is breached when others make inferences about this individual[61]. Jeff Skopek, a law professor at Cambridge University, argues that "data mining often generates knowledge about people through the process of inference rather than direct observation or access, and there are both legal and normative grounds for rejecting the notion that inferences can violate privacy."[62] To put the question another way, consider pregnancy. If a person were to believe that his friend was pregnant by stealing the friend's records from her obstetrician–gynecologist records or by tapping her phone, that would clearly represent a privacy violation. However, if this person reached a belief that his friend was pregnant by seeing that she stopped drinking at dinner, changed her diet, and put on some weight, it is hard to argue that there was a privacy violation. The question is whether big data analysis is more like the former or more like the latter. Of course, big data enables us to make many more inferences with much more confidence than do the friendly observations in the pregnancy example, but is the deontological analysis about the amount we believe we know or the route by which we believe we know it?

A path forward

One reaction to the health privacy violations described above, both deontological and consequentialist, is to sharply limit access to patient data. Particularly if deontological and consequentialist concerns are difficult to decrease ex post, decreasing access to data ex ante seems like an attractive solution[63]. Under this approach, perhaps data sharing should be limited to the minimal amount necessary in all contexts, data should be retained only for limited time, or data should be intentionally obfuscated if consequential

harms are difficult to limit[64]. Nevertheless, we argue that limits on data access can bring their own harms.

The basic harm of privacy overprotection is the brakes it puts on data-driven innovation[65]. Privacy protections limit both data aggregation, whether in the creation of longitudinal records or in the collation of data from different sources at the same time, and innovative data use. As a straightforward example, data deidentification is a common way to comply with HIPAA requirements—but deidentified data are much harder to link together when a patient sees different providers, gets insurance through different payers over time, or moves state-to-state[30,49]. Patchy, fragmented health data make data-driven innovation hard, imposing both technological and economic hurdles.

Some approaches can protect privacy while minimizing the cost to innovation, and these should be pursued. In some contexts, researchers could use techniques involving pseudonymized data or differential privacy rather than identified data[66–68]. Privacy audits can ensure appropriate use and security standards should guard against unauthorized use. Data holders should be stewards of data, not privacy-agnostic intermediaries. But in many contexts, a privacy–innovation tradeoff will still exist.

Privacy also interacts problematically with secrecy. As described above, there are many potential innovations that can arise from data, and some of these may be very lucrative, such as an algorithm that accurately selects cancer drugs. Innovators have incentives to keep data secret to maintain a competitive advantage in development and deployment of such valuable innovations[69]. But we might prefer as a society to have access to the data on which such innovations are based: others can use those data to create better predictors from the same data, to aggregate data to find more subtle patterns, or to validate and verify that the original innovator's research was accurate.

Myriad Genetics' maintenance of a proprietary database of the genetic sequences and medical history of women who sought *BRCA1* and *BRCA2* breast and ovarian cancer predisposition tests exemplifies these concerns; non-Myriad tests returned variants of unknown significance more frequently because Myriad's data were unavailable, and the data could not be aggregated to provide even better tests[70]. Privacy concerns can provide a shield—rhetorical or not—for this type of practice; to the extent that firms can justify keeping proprietary data on the basis that they are protecting patients' privacy, data sharing is harder to demand.

Privacy-justified secrecy can erode trust in already opaque big-data innovations. When big data yields surprising insights about how to provide care, providers and patients need to trust the results to implement them. This already creates challenges when the insights come from explicit analyses of big data; when machine-learning and opaque algorithms are involved, trust may be even harder to engender. To the extent that data and algorithms are kept secret under a potentially disingenuous veil of privacy protection, providers and patients will have even less cause for trust in the results[71]. To be sure, there are many medical processes whose inner workings are shrouded by trade secrecy and very opaque to patients, but the media attention to and newness of big data and AI may make patients particularly nervous about their integration into care.

On the other hand, to the extent that patients concerned about privacy refuse to participate in a data-driven system, those algorithms may not even be developed in the first place. Striking the right balance—protecting privacy so that patients are comfortable providing their data, but not allowing privacy to drive secrecy that reduces validation and trust in the potential benefits arising from those data—will be a tricky challenge for proponents of big data, machine learning, and learning health systems. What is more, the answer will not be uniform. The future of big data privacy will be sensitive to data source, data custodian, and type of data, as well as the importance of data triangulation from multiple sources. But it is important that we not assume privacy maximalism across the board is the way to go. Privacy underprotection and overprotection each create cognizable harms to patients both today and tomorrow.

Received: 27 April 2018; Accepted: 30 October 2018;
Published online: 7 January 2019

References

1. Cohen, I. G., Amarasingham, R., Shah, A., Xie, B. & Lo, B. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Aff.* **33**, 1139–1147 (2014).
2. Executive Office of the President. Big data: seizing opportunities, preserving values. https://bigdatawg.nist.gov/pdf/big_data_privacy_report_may_1_2014.pdf (2014).
3. Hoffman, S. *Electronic Health Records and Medical Big Data* (Cambridge Univ. Press, New York, 2016).
4. Institute of Medicine. Committee on Quality of Health Care in America, the National Academies. *To Err is Human: Building a Safer Health System* (eds. Kohn, L. T., Corrigan, J. M., & Donaldson, M. S.) (National Academies Press, Washington, D.C., 2000).
5. Centers for Medicare and Medicaid Services. Hospital inpatient quality reporting program. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html (2017).
6. Kohane, I. S. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* **12**, 417–428 (2011).
7. Behrman, R. E. et al. Developing the sentinel system—a national resource for evidence development. *N. Engl. J. Med.* **364**, 498–499 (2011).
8. Price, W. N. II Black-box medicine. *Harv. J.L. & Tech.* **28**, 419–467 (2016).
9. Terry, N. P. Appification, AI, & healthcare's new iron triangle. Preprint at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3020784 (2018).
10. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
11. Liu, N. T. et al. Development and validation of a machine learning algorithm and hybrid system to predict the need for life-saving interventions in trauma patients. *Med. Biol. Eng. Comput.* **52**, 193–203 (2014).
12. Avati, A. et al. Improving palliative care with deep learning. Preprint at https://arxiv.org/pdf/1711.06402.pdf (2018).
13. Spector-Bagdady, K. & Shuman, A. Reg-ENT within the learning health system. *Otolaryngol. Head. Neck. Surg.* **158**, 405–406 (2018).
14. Price, W. N. II Regulating black-box medicine. *Mich. L. Rev.* **116**, 421–474 (2017).
15. Institute of Medicine. *The LearningHealthcare System: Workshop Summary* (eds. Olsen, L. A., Aisner, D. & McGinnis, J. M.) (National Academies Press, Washington, D.C., 2007).
16. Faden, R. R. et al. An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. *Hastings Ctr. Rep.* **43**, S16–S27 (2013).
17. Kass, N. E. The research-treatment distinction: a problematic approach for determining which activities should have ethical oversight. *Hastings Ctr. Rep.* **43**, S4–S15 (2013).
18. Raval, M. V., Sakran, J. V., Medbery, R. L., Angelos, P. & Hall, B. L. Distinguishing QI projects from human subjects research: ethical and practical considerations. *Bull. Am. Coll. Surg.* **99**, 21–7 (2014).
19. Miller, F. G. & Emanuel, E. J. Quality-improvement research and informed consent. *N. Engl. J. Med.* **358**, 765–767 (2008).
20. Morreim, H. Research versus innovation: real differences. *Am. J. Bioeth.* **5**, 42–43 (2005).
21. Friedman, C. P., Wong, A. K. & Blumenthal, D. Achieving a nationwide learning health system. *Sci. Translat. Med.* **2**, 57cm29 (2010).
22. Nissenbaum, H. *Privacy in Context: Technology, Policy, and the Integrity of Social Life* (Stanford Univ. Press, Stanford, CA, USA, 2010).
23. Konnoth, C. An expressive theory of privacy intrusions. *Iowa L. Rev.* **102**, 1533–1581 (2017).
24. Terry, N. P. Regulatory disruption and arbitrage in health-care data protection. *Yale J. Health Pol'y L. & Ethics* **17**, 143–208 (2017).
25. Terry, N. P. Existential challenges for healthcare data protection in the United States. *Ethics, Med., & Pub. Health* **3**, 19–27 (2017).
26. Commission Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of such Data, and Repealing Directive, 95/46/EC, 2016 O.J. (L 119) 1, 34 (General Data Protection Regulation). http://www.privacy-regulation.eu/en/article-4-definitions-GDPR.htm (2016).
27. Spector-Bagdady, K., Prince, A. E. R., Yu, J. H. & Appelbaum, P. S. Analysis of state laws on informed consent for clinical genetic testing in the era of genomic sequencing. *Am. J. Med. Genet. C. Semin. Med. Genet.* **178**, 81–88 (2018).
28. 45 C.F.R. §§ 160.103–164.504.
29. 45 C.F.R. §§ 164.302–318.

30. Eisenberg, R. S. & Price, W. N. II Promoting healthcare innovation on the demand side. *J.L. & Biosciences* **4**, 3–49 (2017).
31. 45 C.F.R. § 164.514.
32. Gymrek, M. et al. Identifying personal genomes by surname inference. *Science* **339**, 321–324 (2013).
33. National Committee on Vital and Health Statistics and its Privacy, Security, and Confidentiality Subcommittee, U.S. Department of Health and Humam Services. Health information privacy beyond HIPAA: a 2018 environmental scan of major trends and challenges. https://ncvhs.hhs.gov/wp-content/uploads/2018/05/NCVHS-Beyond-HIPAA_Report-Final-02-08-18.pdf (2017).
34. Philibert, R. A. et al. Methylation array data can simultaneously identify individuals and convey protected health information: an unrecognized ethical concern. *Clin. Epigenetics* **6**, 28 (2014).
35. Centers for Medicare and Medicaid Services. Blue Button® 2.0: improving medicare beneficiary access to their health information. https://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information-Technology/Blue-Button/index.html.
36. Couzin-Frankel, J. After a prominent gene-testing firm declined to give patients their complete data, ACLU filed a complaint. *Science* https://www.sciencemag.org/news/2016/05/after-prominent-gene-testing-firm-declined-give-patients-their-complete-data-aclu-filed (2016).
37. Riley, M. F. Big data, HIPAA, and the common rule: time for a big change? In *Big Data, Health Law, and Bioethics* (eds. Cohen, I. G., Fernandez Lynch, H., Vayena, E. & Gasser, U.) (Cambridge Univ. Press, New York, 2018).
38. Hoffman, S. Citizen science: the law and ethics of public access to medical big data. *Berkeley Tech. L.J.* **30**, 1741–1805 (2015).
39. Barocas, S. & Selbst, A. D. Big data's disparate impact. *Calif. L Rev.* **104**, 671–732 (2016).
40. Malanga, S. E., Loe, J. D., Robertson, C. T. & Ramos, K. S. Who's left out of big data? how big data collection, analysis, and use neglects populations most in need of medical and public health research and interventions. In *Big Data, Health Law, and Bioethics* (eds. Cohen, I. G., Fernandez Lynch, H., Vayena, E. & Gasser, U.) (Cambridge Univ. Press, New York, 2018).
41. Chen, I., Johansson, F. D. & Sontag, D. Why is my classifier discriminatory? Preprint at https://arxiv.org/pdf/1805.12002.pdf (2018).
42. Kleinberg, J., Mullainathan, S. & Raghavan, M. Inherent trade-offs in the fair determination of risk scores. Preprint at https://arxiv.org/pdf/1609.05807.pdf (2016).
43. Cohen, I. G. Is there a duty to share health care data? In *Big Data, Health Law, and Bioethics* (Cohen, I. G., Fernandez Lynch, H., Vayena, E. & Gasser, U. eds., Cambridge Univ. Press, New York, 2018).
44. Kaye, J. et al. Dynamic consent: a patient interface for twenty-first century research networks. *Eur. J. Hum. Genet.* **23**, 141–146 (2015).
45. Grady, C. et al. Broad consent for research with biological samples: workshop conclusions. *Am. J. Bioeth.* **15**, 34–42 (2015).
46. Mayer-Schönberger, V. & Ingelsson, E. Big data and medicine: a big deal? (Review Symposium). *J. Intern. Med.* **283**, 418–429 (2018).
47. Rockhold, F., Nisen, P. & Freeman, A. Data sharing at a crossroads. *N. Engl. J. Med.* **375**, 1115–1117 (2016).
48. Winickoff, D. & Winickoff, M. The charitable trust as a model for genomic biobanks. *N. Engl. J. Med.* **349**, 1180–1184 (2003).
49. Evans, B. J. Big data and individual autonomy in a crowd. In *Big Data, Health Law, and Bioethics* (eds. Cohen, I. G., Fernandez Lynch, H., Vayena, E. & Gasser, U.) (Cambridge Univ. Press, New York, 2018).
50. Maschke, K. J. Governance Issues for Biorepositories and BiospecimenResearch 299. In *Specimen Science: Ethics and Policy Implications* (eds. Lynch, H. F., Bierer, B. E., Cohen, I. G. & Rivera, S. M.) (MIT Press, Cambridge, MA, USA, 2017).
51. Connected Health Cities. *Citizens' Juries Report*. https://www.connectedhealthcities.org/what-is-a-chc/public-engagment/citizens-juries-chc/citizens-juries/ (2017).
52. Calo, M. R. The boundaries of privacy harm. *Indiana L.J.* **86**, 1131–1162 (2011).
53. Epstein, R. A. The legal regulation of genetic discrimination: old responses to new technology. *B.U. L. Rev.* **74**, 1–23 (1994).
54. Stone, D. A. The struggle for the soul of health insurance. *J. Health Polit. Policy & L.* **18**, 287–317 (1993).
55. Hoffman, A. K. Three models of health insurance: the conceptual pluralism of the Patient Protection and Affordable Care Act. *U. Penn. L. Rev.* **159**, 1873–1954 (2011).
56. Hoffman, S. data's new discrimination threats: amending the americans with disabilities act to cover discrimination based on data-driven predictions of future disease. In *Big Data, Health Law, and Bioethics* (eds. Cohen, I. G., Fernandez Lynch, H., Vayena, E. & Gasser, U. eds.) (Cambridge Univ. Press, New York, 2018).
57. Mello, M. M., Lieou, V. & Goodman, S. N. Clinical trial participants' views of the risks and benefits of data sharing. *N. Engl. J. Med.* **378**, 2202–2211 (2018).
58. Grande, D. et al. Public preferences about secondary uses of electronic health information. *JAMA Intern. Med.* **173**, 1798–1806 (2013).
59. Ford, R. A. & Price, W. N. II Privacy and accountability in black-box medicine. *Mich. Telecomm. & Tech. L. Rev.* **23**, 1–43 (2016).
60. May, T. Sociogenetic risks—ancestry DNA testing, third-party identity, and protection of privacy. *N. Engl. J. Med.* **379**, 410–412 (2018).
61. Crawford, K. & Schultz, J. Big data and due process: toward a framework to redress predictive privacy harms. *B.C. L. Rev.* **55**, 93–128 (2014).
62. Skopek, J. M. Big data's epistemology and its implications for precision medicine and privacy. In *Big Data, Health Law, and Bioethics* (eds. Cohen, I. G., Fernandez Lynch, H., Vayena, E. & Gasser, U.) (Cambridge Univ. Press, New York, 2018).
63. Terry, N. P. Protecting patient privacy in the age of big data. *U.M.K.C. L. Rev.* **81**, 1–34 (2012).
64. Goldacre, B. How to get all trials reported: audit, better data, and individual accountability. *PLoS. Med.* **12**, e1001821 (2015).
65. Price II, W. N. Drug approval in a learning health system. Preprint at https://papers.ssrn.com/abstract_id=3152570 (2018).
66. Beaulieu-Jones, B. K. et al. Privacy-preserving generative deep neural networks support clinical data sharing. Preprint at https://www.biorxiv.org/content/early/2018/06/05/159756 (2018).
67. Dwork, C. & Roth, A. The algorithmic foundations of differential privacy. *Found. & Trends in Theoretical Comput. Sci.* **9**, 211–407 (2014).
68. Moussa, M. & Demurjian, S. A. Differential privacy approach for big data privacy in healthcare. In *Privacy and Security Policies in Big Data* (eds. Tamane, S., Solanki, V. K. & Dey, N. eds.) (IGI Global, Hershey, PA, USA, 2017).
69. Price, W. N. II Big data, patents, and the future of medicine. *Cardozo L. Rev.* **37**, 1401–1453 (2016).
70. Cook-Deegan, R. et al. The next controversy in genetic testing: clinical data as trade secrets? *Eur. J. Hum. Genetics* **21**, 585–588 (2013).
71. Spector-Bagdady, K. "The Google of Healthcare:" enabling the privatization of genetic bio/databanking. *Ann. Epidemiol.* **26**, 515–519 (2016).
72. Greely, H. T. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. *Annu. Rev. Genomics Hum. Genet.* **8**, 343–346 (2007).
73. Ohm, P. Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA L. Rev.* **57**, 1738–1777 (2010).
74. Narayanan, A. & Shmatikov, V. Robust deanonymization of large sparse datasets (how to break the anonymity of the Netflix prize database). In *2008 IEEE Symposium on Security and Privacy.* http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf (2008).

Acknowledgements

The authors extend thanks to N. Terry and K. Spector-Bagdady.

Competing interests

W.N.P. and I.G.C.'s research reported in this publication was done with the support of CeBIL (Collaborative Research Program for Biomedical Innovation Law). CeBIL is a scientifically independent collaborative research program supported by a Novo Nordisk Foundation Grant (grant number NNF17SA0027784). W.N.P.'s work was also supported by the National Cancer Institute (Grant number 1-R01-CA-214829-01-A1; The Lifecycle of Health Data: Policies and Practices). I.G.C. has served as a consultant for Otsuka Pharmaceuticals on their Abilify MyCite product.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to I.G.C.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.