VIEWPOINT

Big Data, Big Tech, and Protecting Patient Privacy

I. Glenn Cohen, JD
Harvard Law School and the Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics, Harvard University, Cambridge, Massachusetts.

Michelle M. Mello, JD, PhD
Stanford Law School and Stanford University School of Medicine, Stanford, California.

The market for patient data has never been more active. Technology companies, from startups to giants, are eager to access electronic health record (EHR) data to build the next generation of health-focused products. Medical artificial intelligence (AI) is particularly data-hungry; large, representative data sets hold promise for advancing not only AI companies' growth, but also the health of patients.[1] Companies' overtures to major hospitals about data sharing have highlighted legal and ethical uncertainties as to whether and how to undertake these relationships.

One such partnership is now being challenged in court. In June 2019, a patient sued the University of Chicago Medical Center and Google for alleged misuse of patient EHR data.[2] This Viewpoint discusses the case and what it signals about the need for thoughtful governance of data sharing between health care organizations and technology companies.

The Complaint

In *Dinerstein v Google*, a class action complaint filed in federal court in Illinois, Matt Dinerstein asserted that the University of Chicago violated his privacy by turning over his and thousands of other patients' EHR data to Google. The forms Dinerstein signed at the hospital stated that his medical records would not be disclosed to third parties for commercial purposes. Although most identifying information was removed, the records given to Google allegedly contained date stamps indicating the dates and times that services were rendered, as well as free-text notes from clinical visits. The complaint, which seeks money damages and a court order to stop the use and further transfer of patient records, alleges deceptive and unfair business practices in violation of state consumer protection law, breach of contract, violation of common-law privacy rights, and other claims.

Google's interest in medical records derived from its plans to develop a novel EHR system. It envisions a system that uses AI to analyze patient records, predict future clinical events, and highlight what past medical information contributed to the predictions (**Figure**). Google aspires to reduce information overload and help clinicians decide "which patients have the highest need for my attention now and, at an individual level, what information in the patient's chart should I attend to?"[3]

HIPAA Shows Its Age

Releasing date stamps (an allegation the University of Chicago denies) would constitute a clear violation of the Health Insurance Portability and Accountability Act (HIPAA), the flagship health information privacy law in the United States. HIPAA requires patient authorization (or a decision by an institutional review or privacy board to waive that requirement) to disclose dates of care and other specified, potentially identifying, data fields. But the complaint is not brought under HIPAA; rather, it uses the alleged violations as evidence for other claims. Although the possibility of a massive payout to resolve HIPAA violations is a major concern of hospital attorneys, the statute is enforceable by federal agency action only. That is not the only shortcoming of HIPAA as a vehicle for redressing the types of harms at issue in *Dinerstein v Google*. HIPAA is a 20th-century statute ill equipped to address 21st-century data practices. When HIPAA was adopted in 1996, the internet had 20 million US users who spent an average of half an hour per month browsing. Google did not exist, the global internet had about 100 000 websites, and geolocation tracking was available only for the military. Few health care organizations and clinicians had adopted EHRs. Personal data were presumed nonidentifiable if stripped of 18 identifiers, most of which were direct identifiers (eg, phone numbers).

The world described in the *Dinerstein v Google* case is radically different. EHRs are highly penetrant and often interlinked; search engines track user activity on nearly 2 billion websites, many of which collect additional information; smartphones permit detailed geolocation tracking; and an entire industry of data aggregators pools and packages consumer information for analysis and resale. The substantial increase in available personal information and advances in computing mean that individuals can often be identified in deidentified data by triangulating data sources. The *Dinerstein v Google* complaint recounts in detail how Google could, in theory, combine patients' geolocation and other smartphone data with dates and times in the EHR to determine who visited which clinical departments when, which services they received, and what notes their physicians wrote about them. Once an individual's identity is ascertained, the company could then link EHR data with other types of information about that person (eg, what they purchase).

HIPAA bars none of this except the release of date stamps, and would not be implicated, for example, if Google identified individuals by linking EHR data without HIPAA identifiers to internet data of consumers who visited the University of Chicago hospital and searched online for information about particular medical conditions or if a social media company linked such EHR data to a user's posts about a hospital stay. Such concerns have prompted the widespread suggestion that it is time for a privacy law reboot in the United States, as there has been in Europe.

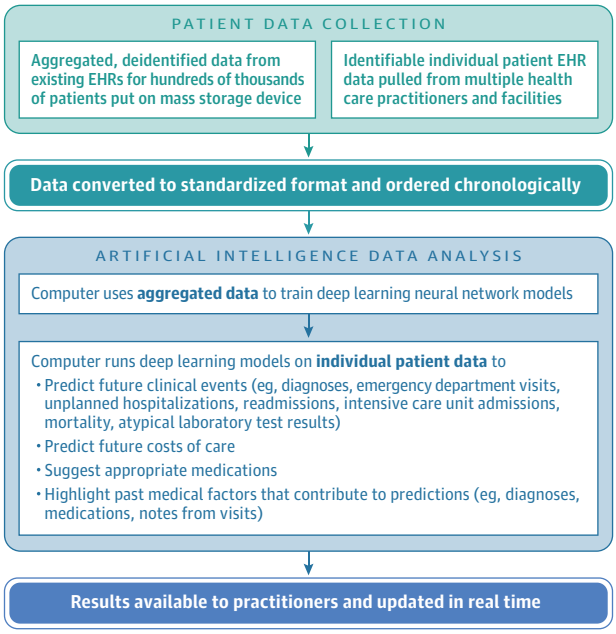
A Better Approach?

As the growing promise of AI to improve health care butts up against obsolescent privacy laws and public distrust of powerful, lightly regulated "big tech" companies, cases like *Dinerstein v Google* seem inevitable. The case surfaces important questions about whether to further regulate patient data sharing, who should regulate it, and how it should be regulated.

Arresting the use of (putatively) deidentified patient data in product development seems neither feasible nor

Corresponding Author: I. Glenn Cohen, JD, Harvard Law School, Griswold Hall, Room 503, Cambridge, MA 02138 (igcohen@law.harvard.edu).

Figure. Potential Next-Generation Electronic Health Record (EHR) System Built Using Deidentified Patient Records



Adapted from the patent application by Mossing et al.[3]

desirable, and problems with current EHR systems provide insights for understanding why. Today's clunky EHR systems are reviled by clinicians and known contributors to professional frustration, often characterized as burnout.[4] Many physicians would be excited to see Google disrupt the market. Yet, some patients have serious discomfort with corporations using their health information (especially when large profits are involved), and deidentification is not the tonic it used to be. Patient concerns could be addressed by clinicians seeking patient authorization for sharing even deidentified information outside the patient care operation. Patients could be presented with a blanket "front door" authorization form and choose to sign or withhold permission. However, this approach may prove to be mere ethical window dressing.[5,6] HIPAA appropriately calls such a process *authorization*, not *consent*, because patients are rarely given the information and opportunity to ask questions needed to give meaningful informed consent to future uses of their data. Even if those problems could be overcome, it is asking a great deal of patients to imagine and assess how their information may be used and what the risk of reidentification may be. Further, what happens when the benefits or risks of EHR data sharing change over time? It may not be feasible to withdraw deidentified data that have already been shared. In addition, opt-outs to data sharing are unlikely to be randomly distributed. Evidence suggests some racial and ethnic minorities are especially likely to have concerns about blanket use of their data, which may further enervate their representation in data sets used to build models that affect their care.[1]

Some privacy scholars favor reframing authorization as one possible "governance technology" among many.[1,5,6] Authorization that is individualized, upstream (ie, obtained early), and typically one-and-done can be supplemented with governance that is group-based, downstream (ie, obtained at the time of particular uses), and ongoing. In this approach, a committee would review requests for specific uses of deidentified patient data. Analogs include existing committees that evaluate compassionate-use requests for pharmaceuticals and requests to access repositories of biospecimens, participant-level clinical trial data, and government-held demographic data and an initiative to create an independent board to decide when researchers may use Facebook data.[7] Proposals for such data access committees typically envision an entity that makes decisions independent of the influence of data holders or data users but that may sit within either type of organization. Its members would include statisticians and programmers to evaluate reidentification risks and experts in ethics and particular types of research, but at least half of the members should be patients of the institution whose EHR data are being sought. The patients would receive significant training and be expected to make important intellectual contributions to decision-making. Permission for data uses would be purpose-specific and contingent upon the user's fulfillment of ongoing reporting obligations.

Rather than relying on individual patients to make decisions about variable, dynamically evolving risks and benefits of potential uses of their data, this approach could enable a group of patients to be properly trained to make informed, specific decisions. Deliberations can nimbly change course as new facts emerge. The most important weakness is one common to all representational approaches: how well the views of the representatives reflect those of the people they represent. Despite this problem, this approach is preferable to both the current situation and a more extensive authorization-based regime. Even if not required by future regulation, hospitals sharing deidentified EHR data with third parties would be well advised to put such structures in place for ethical (and, cynically, reputational) reasons.

Conclusions

Dinerstein v Google exposes a US health data regulatory regime that is showing its age. Litigation may foster improvements at the margins, but what is really needed is a thorough rethinking of data sharing governance for the 21st century.

ARTICLE INFORMATION

Published Online: August 9, 2019. doi:10.1001/jama.2019.11365

Conflict of Interest Disclosures: Mr Cohen reported serving as a consultant for Otsuka Pharmaceuticals on Abilify MyCite and receiving a grant from the Collaborative Research Program for Biomedical Innovation Law, a scientifically independent collaborative research program supported by a Novo Nordisk Foundation grant (NNF17SA0027784). Dr Mello reported receiving grants from the Greenwall Foundation, a nonprofit foundation focused on bioethics, for work in the topic area of the article.

REFERENCES

1. Price WN II, Cohen IG. Privacy in the age of medical big data. *Nat Med*. 2019;25(1):37-43.

2. Complaint, *Dinerstein v Google*, No. 1:19-cv-04311 (Ill 2019).

3. Mossing A, Rajkomar A, Oren E, et al, inventors; Google Inc, assignee. System and method for predicting and summarizing medical events from electronic health records. US patent application 62/538,112. July 28, 2017.

4. Tai-Seale M, Dillon EC, Yang Y, et al Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. *Health Aff (Millwood)*. 2019;38(7):1073-1078.

5. Parasidis E, Pike E, McGraw D. A Belmont report for health data. *N Engl J Med*. 2019;380(16):1493-1495. doi:10.1056/NEJMp1816373

6. Vayena E, Blasimme A. Health research with Big Data. *J Law Med Ethics*. 2018;46(1):119-129.

7. King G, Persily N. A new model for industry-academic partnerships. Gary King website. https://gking.harvard.edu/files/gking/files/partnerships.pdf. Updated May 29, 2019. Accessed July 31, 2019.