3-7-2017

Risk and Resilience in Health Data Infrastructure

Nicholson Price
University of Michigan Law School, wnp@umich.edu

RISK AND RESILIENCE IN HEALTH DATA INFRASTRUCTURE

W. Nicholson Price II, PhD [*]

Today's health system runs on data. However, for a system that generates and requires so much data, the health care system is surprisingly bad at maintaining, connecting, and using those data. In the easy cases of coordinated care and stationary patients, the system works—sometimes. But when care is fragmented, fragmented data often result.

Fragmented data create risks both to individual patients and to the system. For patients, fragmentation creates risks in care based on incomplete or incorrect information, and may also lead to privacy risks from a patched-together system. For the system, data fragmentation hinders efforts to improve efficiency and quality, and to drive health innovation based on collected data.

Efforts to combat data fragmentation would benefit by considering the idea of health data infrastructure. Most obviously, that would be infrastructure for health data—that is, infrastructure on which health data can be stored and transmitted. But it should also be an infrastructure of health data—that is, a platform of shared data on which to base further efforts to increase the efficiency or quality of care.

Today's health system runs on data. Patients and doctors complain about the proportion of time during a patient appointment that is spent entering data into the doctor's computer, but this has become the new normal. Data are supposed to help improve care for individual patients, to increase the efficiency of the system as a whole, and to provide the basis for future innovation in care.

However, for a system that generates and requires so much data, the health care system is surprisingly bad at maintaining, connecting, and using those data. In the easy cases, it works. If a patient stays with the same primary care physician, coordinates all care through that physician, goes to the same pharmacy, the same hospital, and the same labs, and uses the same insurer, that patient's records may—*may*—be integrated into a single comprehensive medical record that tracks the patient's health over time. But patients don't behave like this most of the time. Patients move between providers, pick up

---

drugs while traveling, switch insurers as they change jobs (or lose them), see different specialists, and generally vary the parameters of their care. And the health data system does a poor job accounting for this fragmentation of care, resulting in fragmented data.

Fragmented data create risks to patients and to the system as a whole. At the patient level, fragmentation creates risks in care, where information necessary for effective care is either not available or incorrect. Fragmentation also creates risks for patient privacy, as a result of the needs to haphazardly share data across different health actors. At the systemic level, data fragmentation hinders efforts to make the system more efficient as a whole, because putative optimizers only see a fragment of the picture. It also slows innovation in health, especially big-data driven modern initiatives that rely on large, high-quality datasets for their power and accuracy.

Efforts to combat data fragmentation would benefit by considering the idea of health data infrastructure. Most obviously, that would be infrastructure *for* health data—that is, infrastructure on which health data can be stored and transmitted. But it should also be an infrastructure *of* health data—that is, a platform of shared data on which to base further efforts to increase the efficiency or quality of care.

This essay proceeds in three Parts. Part I describes the landscape of health data today, including potential benefits of the collection of health data and the reasons for fragmentation which limits those benefits. Part II describes the risks of a fragmented health data system. Part III sketches the basics of how an infrastructure vision for and of health data might look.

I. HEALTH DATA TODAY

The health system generates a blizzard of data at an increasing rate. From the paper records of prior practice, providers have largely moved to use electronic health records (also called electronic medical records).[1] New forms of data are proliferating to fill those records, including the reports of traditional medical encounters, high-volume diagnostic tests such as genetic sequencing and analysis, prescription records, and others.[2]

---

[1] The move to electronic health records was not accidental. A substantial sum was made available for providers to shift to electronic records HITECH Act, passed as part of the American Recovery and Reinvestment Act of 2009, Pub. L. No. 111-5, 123 Stat. 115 (Feb. 17, 2009) (ARRA), Div. A, Title XIII, Div. B, Title IV. *See* SHARONA HOFFMAN, ELECTRONIC HEALTH RECORDS & MEDICAL BIG DATA 38–40 (2016). As a powerful counterpart, penalties are imposed on entities failing to shift to and meaningfully use electronic records by established deadlines. *See id.* at 41–42; Centers for Medicare and Medicaid Services, *Medicare and Medicaid EHR Incentive Program Basics*, Jan. 12, 2016, https://www.cms.gov/regulations-and-guidance/legislation/ehrincentiveprograms/basics.html.

[2] *See* Rebecca S. Eisenberg & W. Nicholson Price II, *Promoting Health Innovation on the Demand Side*, __ J.L. & BIOSCIENCES __ (2017).

A. Potential Benefits

These data are collected for a reason; they are supposed to create substantial benefits for patients, providers, and for the health system as a whole. Ideally, they should lead to improved care for individual patients as integrated medical records prevent easily avoidable medical error and allow a broader picture of the patient's overall health.[3] They should enable more efficient care by reducing the costs of coordination, should decrease costs, and should even enable more effective and efficient billing by insurers. On a slightly more systemic level, many health care reforms rely on the ability to measure care precisely—for instance, to observe whether patients are treated according to approved procedures or are readmitted to hospitals too frequently.[4] Health data enable the imposition of sanctions or the provision of incentives to try to shape health care in productive ways.[5]

Data are also supposed to enable us to draw more nuanced and useful information from the health system. Insurers and others have used information about actual patient experience in the health system to demonstrate that certain drugs are less safe than expected,[6] that some treatments may be more cost-effective at providing the same benefit,[7] that some patients gain more benefit from a particular treatment than others,[8] or that a drug should be moved from prescription-only to over-the-counter status.[9] Recently, FDA has even gained the statutory authority to use this type of real-world evidence to approve new indications for drugs.[10] More broadly,

---

[3] *See, e.g.*, James R Broughman & Ronald C Chen, *Using Big Data for Quality Assessment in Oncology*, 5 J. COMP. EFF. RES. 309 (2016).

[4] *See, e.g.*, Broughman & Chen, *supra* note 3.

[5] *See* Medicare Access and CHIP Reauthorization Act of 2015 (MACRA), Pub. L. No. 114-10, 129 Stat. 87, § 102 (requiring a plan to develop data-based measures for physician and hospital performance), § 101 (creating payment incentive structures using those measures).

[6] *See* Eisenberg & Price, *supra* note 2, at __ (discussing the identification of toxic side effects of the painkiller Vioxx by Kaiser Permanente, which analyzed patient records in its integrated health system and found higher rates of heart attacks among patients taking Vioxx than among patients taking other similar drugs).

[7] *See id.* § I.C.2 (describing cost-effectiveness research and the use of observational studies of patient data to perform such research).

[8] *See id.* (describing comparative-effectiveness research).

[9] *Id.* at § I.A.1 (describing a petition filed by Blue Cross of California (later Wellpoint) to take certain antihistamines, including Claritin, over-the-counter).

[10] *See* 21st Century CURES Act, Pub. L. No. 114-255, § 3022 (requiring FDA to "establish a program to evaluate the potential use of real world evidence" for the approval of new indications for an already-approved drug or to fulfill post-approval study or surveillance requirements). This provision has been the subject of considerable criticism. *See, e.g.*, Jerry Avorn & Aaron S. Kesselheim, *The 21st Century Cures Act — Will It Take Us Back in Time?*, 372

health data can potentially lead to advances in precision medicine. Precision medicine, the scientific tailoring of medical treatment to reflect individual patient variation, requires knowing how different patients respond to different forms of treatment.[11] Some of this knowledge can be generated by classical hypothesis-driven scientific and clinical studies, but other advances, including those relying on machine-learning and other forms of datamining, rely on large sets of existing health data.[12]

Overall, health data offer substantial promise for improving health care, both in terms of near-term patient-specific benefits and in terms of later innovations to improve the health system. Unfortunately, these benefits have been slow to materialize. At least in part, this slowness has resulted from the fragmentation of health data.[13]

B. Fragmentation

Why are health data today so fragmented? There are at least three linked reasons. First, and most obviously, care itself is fragmented. Second, and related, competition between entities in the health system reduces incentives to connect and link data. Third and finally, legal barriers to information sharing, especially the Health Insurance Portability and Accountability Act, make it hard to link data.

1. Fragmented care

The key underlying cause of health data fragmentation is that health care is itself fragmented, and with it the generation and storage of health data.[14]

N. ENGL. J. MED. 2473 (2015).

[11] Laura K. Wiley et al., *Harnessing next-Generation Informatics for Personalizing Medicine: A Report from AMIA's 2014 Health Policy Invitational Meeting*, 23 J. AM. MED. INFORM. ASSOC. 413 (2016); Marc L Berger et al., *Opportunities and Challenges in Leveraging Electronic Health Record Data in Oncology*, 12 FUTURE ONCOL. 1261 (2016).

[12] *See* W. Nicholson Price II, *Black-Box Medicine,* 28 HARV. J.L. & TECH. 419, 429–34, 437–39 (2015) (describing the big data potential and requirements of next-generation black-box medicine).

[13] The fragmentation of health data is certainly not the only cause for the delay in realizing benefits of health data innovation. Some actors lack the right incentives to actively move toward the highest-quality, most efficient care. *See, e.g.*, Eisenberg & Price, *supra* note 2, at __ (discussing the problematic incentives for health insurers and for drug manufacturers); David Orentlicher, *Paying Physicians More to Do Less: Financial Incentives to Limit Care*, 30 UNIV. RICHMOND L. REV. 155 (1996) (discussing the incentives of doctors to provide more care than necessary). Technological hurdles also play a role. *See* Eisenberg & Price, *supra* note 2, at § I.D. And even once innovative information is generated, getting health care providers to implement the new knowledge can be challenging. *Id.* at § II.B.

[14] *See, e.g.*, Alan M. Garber & Jonathan Skinner, *Is American Health Care Uniquely Inefficient*, 22 J. ECON. PERSP. 27 (2008) (noting popular wisdom that the American health care system is exceptionally fragmented).

Patients see different doctors at different times, visit different drugstores, change insurers, and in other ways participate in an inherently fragmented health system.[15] Correspondingly, hospitals, doctors, insurers, and pharmacies all keep their own records. These records are generated for different purposes and may use different terms or code different information.[16] For instance, insurance claims records are principally generated for the purpose of payment; accordingly, they lack some forms of care data and may potentially be skewed.[17] The relevant information about patient care is thus spread among different actors in the health care system, in different forms.

Health data are not only generated in the course of health care. Research companies like 23andMe collect substantial health information[18] but are not involved in care, and keep their data separate—potentially to be used for later commercial research. Non-care entities, like Fitbit (whose activity trackers monitor physical activity),[19] Apple (which aims to create a personal digital hub of health information),[20] or others, also generate health data—but they are, of course, largely separate from the system of health and hold different data in different places as well. Overall, different entities both within and outside the health care system generate data separately, which are then held in different silos. This might not be so problematic if communication and data-sharing between the silos were easy and seamless. Unfortunately, it isn't.

2. Data competition

Even for parallel entities, like multiple doctors that a patient may see, competition also keeps data fragmented. Theoretically, among care providers, competition should be irrelevant; the duty of care to patients should preclude competitive hoarding of data or refusal to share data. But no such pressure exists for the providers of diagnostic tests, for instance, or among others that

---

[15] *See* Eisenberg & Price, *supra* note 2, at § II.B.

[16] *Id.*

[17] *Id.* at I.D.

[18] Antonio Regalado, *23andMe Sells Data for Drug Search*, MIT TECH. REV. (June 21, 2016), https://www.technologyreview.com/s/601506/23andme-sells-data-for-drug-search/ (describing 23andMe's collection of data and its sales of data subsets to over a dozen drug companies, including to Genentech for $10 million to search for Parkinson's drugs).

[19] Other sports companies are getting into the health data game. For instance, Nike recently signed a multimillion-dollar deal to collect and analyze performance data collected from athletes at the University of Michigan. Marc Tracy, *With Wearable Tech Deals, New Player Data Is Up for Grabs*, N.Y. TIMES (Sep. 9, 2016), http://www.nytimes.com/2016/09/11/sports/ncaafootball/wearable-technology-nike-privacy-college-football.html.

[20] *See* Apple, *iOS-Health*, http://www.apple.com/ios/health/ (describing the iOS Health App, which collects phone data and can serve as a repository for personal medical records).

collect health or health-related data.[21]

In addition to competition between those who generate data, there is competition between the vendors who provide ways of generating and managing data. The electronic health record market is itself fragmented, with hundreds of vendors.[22] This itself could lead organically to fragmentation through interoperability, as different vendors develop and sell different systems that might happen not to work with each other. However, there is evidence that electronic health record vendors do more, deliberately designing systems that are mutually incompatible to lock customers in and prevent easy migration between systems.[23] This lack of interoperability obviously hinders consolidation of data, transfers between providers as patients move, and the integration of care.

3. Legal barriers

A third barrier to integrating health data comes from legal barriers to data-sharing, especially the Health Insurance Portability and Accountability Act, commonly known as HIPAA.[24] HIPAA places limits on how personally identifiable health data may be used and disclosed.[25] In general, all uses and disclosures of such information by covered entities—providers, insurers, and health data clearinghouses[26]—are prohibited unless specifically permitted. To

---

[21] Perhaps the most well-documented such proprietary data silo is that held by Myriad Genetics, which amassed a dataset of information about women tested for mutations in the breast-cancer-related BRCA1 and BRCA2 genes while it held patents on those genes. *See, e.g.,* Misha Angrist & Robert Cook-Deegan, *Distributing the Future: The Weak Justifications for Keeping Human Genomic Databases Secret and the Challenges and Opportunities in Reverse Engineering Them*, 3 APPL. TRANSL. GENOMICS 124 (2014) (describing Myriad's dataset and others like it); Dan L. Burk, *Patents as Data Aggregators in Personalized Medicine*, 21 BU J. SCI. & TECH. L. 233 (2015) (describing how patents led to Myriad's competitive advantage).

[22] *See* OFFICE OF THE NATIONAL COORDINATOR FOR HEALTH INFORMATION TECHNOLOGY, HOSPITAL HER VENDORS (July 2016), https://dashboard.healthit.gov/ quickstats/pages/FIG-Vendors-of-EHRs-to-Participating-Hospitals.php. The top six vendors provide services for 92% of all nonfederal acute-care hospitals. *Id.*

[23] *See* OFFICE OF THE NATIONAL COORDINATOR FOR HEALTH INFORMATION TECHNOLOGY, REPORT TO CONGRESS: REPORT ON HEALTH INFORMATION BLOCKING 11– 19 (April 2015), www.healthit.gov/sites/default/files/reports/info_blocking_040915.pdf (defining "information blocking" as "when persons or entities knowingly and unreasonably interfere with the exchange or use of electronic health information" and providing evidence of such practices).

[24] Pub. L. No. 104-191, 100 Stat. 2548.

[25] HIPAA's principal data restrictions come from the Privacy Rule, codified at 45 C.F.R. §§ 150ff. HIPAA's regulatory structure is complex and need not be discussed in full here; for additional information, *see, e.g.,* U.S. Dept. of Health and Human Services, *Summary of the HIPAA Privacy Rule* (May 2003), https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/ (providing HIPAA overview); Eisenberg & Price, *supra* note 2, at __ (discussing the Privacy Rule in the context of research using existing health data).

[26] 45 C.F.R. § 160.103. Uses or disclosures by the business associates of covered entities

---

be sure, some permissions are quite broad, such as the use or disclosure of information for the purpose of "health care operations." Theoretically, this should make it easy to share information related to patient care. But HIPAA still creates substantial informal barriers; providers and insurers are notorious for refusing to share information with the blanket invocation of HIPAA, including for uses expressly permitted.[27]

HIPAA creates more substantial and formal barriers to sharing information for secondary research purposes. Research is expressly *not* a permitted purpose for use or disclosure of protected health information.[28] As a result, secondary research often involves health information that has been de-identified, which takes it out of HIPAA's ambit.[29] However, as I have discussed elsewhere, de-identification can increase the fragmentation of health data, because reassembling data about a patient from different sources becomes substantially more difficult—deliberately so—without identifying information.[30] Finally, HIPAA creates barriers between different types of entities that assemble or create health data. HIPAA governs only "covered entities" that are directly involved in the health system. But increasingly, relevant health information is held by entities outside the that system, such as 23andMe, Fitbit, Apple, or others. None of these entities, or the data they hold, are directly governed by HIPAA.[31] Setting aside concerns this raises about fragmented *governance* of health data,[32] it also helps encourage fragmentation through disparate treatment of different entities with different forms of health data.

Notably, there have also been governmental efforts to encourage

---

are governed, though by contract rather than directly under HIPAA's Privacy Rule. 45 C.F.R. § 152(a)(3).

[27] For examples of refusals to share information, *see*, *e.g.*, Paula Span, *Hipaa's Use as Code of Silence Often Misinterprets the Law*, N.Y. TIMES (July 17, 2015), http:// www.nytimes.com/ 2015/07/21/health/hipaas-use-as-code-of-silence-often-misinterprets-the-law.html?_r=0.

[28] 21 C.F.R. § 164.501. Notably, an initial version of the 21st Century CURES Act included a provision adding research as a permissible purpose for use or, directing the Secretary of Health and Human Services to "revise or clarify" the Privacy Rule so that research "including studies whose purpose is to obtain generalizable knowledge" is included as part of the exception for health care operations. *See* H.R. 6 (2015), 114th Congress, § 1124, available at https://www.congress.gov/114/bills/hr6/BILLS-114hr6ih.xml). As passed, the legislation calls instead for the study of such an amendment to the Privacy Rule. Pub. L. No. 114-255 (2016), § 2063.

[29] HIPAA governs only personally identifiable health information; a safe harbor exempts any information from which 17 pieces of identifying information have been removed.

[30] *See* Price, *Patents, Big Data, and the Future of Medicine*, 37 CARDOZO L. REV. 1401, 1413 (2016).

[31] If these entities are business associates of covered entities, they may be regulatd by HIPAA as described in note 26, *supra.*

[32] *See* Nicolas Terry, *Regulatory Disruption and Arbitrage in Healthcare Data Protection*, 17 YALE J. HEALTH POL'Y, L., & ETHICS __ (2017).

interoperability between different health data systems. The Office of the National Coordinator has set out a goal of electronic health record interoperability by 2021 to 2024.[33] And, of course, the push toward electronic health records was itself a federal initiative.[34] Other private systems have been created with the goal of collecting data across providers with the goal of ensuring continuous care and easing the processing of claims; however, these efforts have met with real challenges.[35] Overall, health data in the US health care system remain highly fragmented among different entities, working with different and often mutually incompatible health records systems.

II. Risks of the current system

The risks from a fragmented health data system are substantial. These risks come in two main buckets: primary risks, which is to say risks to patients seeking care in the health system; and secondary risks, which is to say risks that arise when health data are repurposed and used to innovate or improve the system. The primary risks from a fragmented system of health data include, among others, problems in patient care and privacy risks to patient information.

The risks that arise in patient care mirror the potential benefits of electronic health records. If doctors are used to patient information being present in files—to indicate, for example, the presence of an allergy or a drug with potential negative interactions—doctors may be less likely to seek out or independently confirm that information in the absence of an EHR record. This works fine if the information is actually present, but decreases the likelihood of catch an error when the information is missing due to

---

[33] Office of the National Coordinator for Health Information Technology, Connecting Health and Care for the Nation: A 10-Year Vision to Achieve an Interoperable Health IT Infrastructure (2014), http://www.healthit.gov/sites/default/files/ONC10yearInteroperabilityConceptPaper.pdf; Office of the National Coordinator for Health Information Technology, Connecting Health and Care for the Nation: A Shared Nationwide Interoperability Roadmap (Draft Version 1.0 April 2015), http://www.healthit.gov/sites/default/files/nationwide-interoperability-roadmap-draft-version-1.0.pdf

[34] *See* ARRA, *supra* note 1.

[35] For instance, a group of large insurers in California created Cal INDEX, a health information exchange with the goal of automatically collecting and linking patient data from many providers. *See* Cal INDEX, *New California Not-for-Profit to Operate Statewide, Next-Generation Health Information Exchange* (August 5, 2014), https://www.calindex.org/new-california-healthcare-exchange/ (last accessed July 16, 2016) ("Cal INDEX will securely collect and integrate clinical data from providers and claims data from payers to create comprehensive, retrievable patient-centered records known as longitudinal patient records (LPRs)"). The effort has met with limited success thus far. See Beth Kutcher, *Insurers build broad data exchange in California, but providers are slow to join*, Modern Healthcare (March 6, 2016), http://www.modernhealthcare.com/article/20160305/MAGAZINE/303059948.

fragmentation or otherwise.

Similarly, to the extent that failures of interoperability and mistakes from assembling fragmented data introduce active errors in the system, this creates the chance for medical errors which can result in real harm to the patient. If, for instance, a medical administrator receives the records from a previous physician by fax and then adds them by hand to a patient's current record, he might accidentally introduce errors that can compromise future care.[36]

Lastly, when health data aren't meaningfully collected, we lose the opportunity to experience *better*, data-driven care than what we now receive. This isn't a classic "risk," but it does result in costs to patients measured in benefits foregone. To take a simple example, suppose that, as part of a research study, a young woman has her genome sequenced;[37] further suppose that, although this woman not in a high-risk demographic group, she is in fact positive for an allele of the BRCA1 gene that substantially increases her risk of breast cancer. The researcher may not provide her with this information,[38] and there is a substantial likelihood that her genome sequence may be totally separate from her medical records used for primary care. Thus, the patient may not be more rigorously screened for breast cancer, as she would be if had been identified (by that doctor or another involved in her direct care) as a woman with a deleterious BRCA1 allele. In one sense, no new risk has been introduced—but in another, an opportunity for improved care has been missed.

The currently fragmented health data system also creates risks to patient privacy. Patient health data are considered by many to be especially sensitive, meaning that disclosure of such information is an especially substantial privacy concern.[39] Different actors in the system store information in different ways,

---

[36] Sharona Hoffman & Andy Podgurski, *The Use and Misuse of Biomedical Data: Is Bigger Really Better?*, 39 AM. J. LAW MED. 497 (2013).

[37] For the sake of the example, let us assume the lab is CLIA-certified, and that the genetic sequencing is thus of high-enough quality to guide clinical care.

[38] A substantial literature considers the question of returning results from genetic research. For an introduction, *see* Susan M. Wolf et al., *The Law of Incidental Findings in Human Subjects Research: Establishing Researchers' Duties*, 36 J. LAW. MED. ETHICS 361 (2008) (surveying the field); *see also* Ellen Wright Clayton & Amy L. McGuire, *The Legal Risks of Returning Results of Genomics Research*, 14 GENET. MED. 473 (2012) (noting legal risks); R. C. Green *et al.*, *ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing,"* 15 GENETICS MED. 565 (2013) (recommending that a set of identified mutations always be returned to patients); Paul S. Appelbaum et al., *Models of Consent to Return of Incidental Findings in Genomic Research*, 44 HASTINGS CTR. REP. 23 (2014) (noting different models of returning data and different possibilities for informed consent).

[39] *See* Roger A. Ford & W. Nicholson Price II*, Balancing Privacy and Accountability for Black-Box Medicine*, 23 MICH. TELECOMM. & TECH. L. REV. 1 (2017) (describing the privacy concerns related to patient health information); Nicolas Terry, *Protecting Patient Privacy in an Era of Big Data*, 81 UMKC L. REV. 385 (2012).

leading both to less-secure implementations (in, for instance, the office of the solo practitioner that needs to duplicate and keep unnecessary information because it is not available from labs, insurers or specialists directly), and to potential vulnerabilities during information-sharing, when that occurs. Perhaps more importantly, the clunkiness of the system leads to workarounds and kludges that pose inherent security risks. For instance, problems with interoperability (and potentially with HIPAA) may be related to the otherwise-baffling persistence of faxed requests for information between different providers. Hand-answered, unvalidated, and difficult-to-audit fax requests suffer by comparison with high-security, auditable electronic data transfers, but remain the transfer mechanism of choice for some.[40]

The secondary risks from fragmented data come from efforts to use those data for future innovation.[41] Such efforts include the FDA's Sentinel initiative to monitor drug usage for safety risks,[42] observational studies to drive care (which can potentially be used to approve new drug indications under the 21st Century Cures Act[43]), machine-learning efforts to suss out new biological relationships,[44] and implementations of a learning health-care system generally.[45] All of these require that data be high-quality and function much better without substantial gaps in data from different sources or time periods. Fragmentation and errors in health data hinder these efforts. If they don't happen, that is one cost—the foregone benefit of innovation lost. But other risks materialize when innovation relies on incomplete or faulty data. To the extent that new care innovations are based on bad data, they may incorporate errors, biases, or other problems.[46] A fundamental datamining principle is "garbage in, garbage out;" when health care fragmentation creates inaccuracies in data later used in innovation, that innovation suffers, and so may future patients.

---

[40] For instance, the University of Michigan Health System's request for records from another doctor—which itself must be filled out by the patient for each other provider, since no centralized system exists) offers options only for phoning or faxing to request records from another provider.

[41] *See generally* Eisenberg & Price, *supra* note 2 (describing potential innovation by health-care payers using existing health data).

[42] Susan Forrow et al., *The Organizational Structure and Governing Principles of the Food and Drug Administration's Mini-Sentinel Pilot Program*, 21 Pharmacoepidemiol. Drug Saf. 12 (2012).

[43] 21st Century Cures Act, Pub. L. No. 114-255, § 3022 (requiring the Secretary of Health and Human Services to "establish a program to evaluate the potential use of real world evidence . . . to help support the approval of a new indication for a drug.").

[44] *See* Price, *Black-Box Medicine*, *supra* note 12.

[45] *See, e.g.*, Harlan M. Krumholz, *Big Data and New Knowledge in Medicine: The Thinking, Training, and Tools Needed for a Learning Health System*, 33 Health Aff. 1163 (2014).

[46] *See, e.g.*, Sharona Hoffman & Andy Podgurski, *Big Bad Data: Law, Public Health, and Biomedical Databases*, 41 J.L. Med. Ethics 56 (2013).

III. BENEFITS OF RESILIENT HEALTH DATA INFRASTRUCTURE

The risks of fragmented and insecure health data may be at least partially addressed by considering the system in terms of infrastructure—both *for* health data, and *of* health data.

First, the continued fragmentation of health data suggests that the current system is unsustainable. Each actor is responsible for generating, collecting, and storing the data for its own interactions with patients in the health system, and this has led to the substantial risks described above. Given the potential benefits of integrated patient data, effort must be expended at a systemic level to create infrastructure for the sharing, integration, and storage of patient data. This effort need not take any specific form, but the idea of infrastructure for health data, and the risks of fragmented health data, suggest some features of the desired state.

An infrastructure for health data could follow different models of varying centralization. It could exist as a fully centralized health database, where each patient has a single integrated patient record to which different care providers or other entities add data. Alternately, health data could reside in decentralized repositories, much like the current system, but with increased connectivity between the repositories and more rigorous standards that let data be meaningfully transferred between and collated across repositories.[47] This model is closest to the current system—but that closeness demonstrates potential problems, since even with federal initiatives to drive interoperability, fragmentation persists.[48] A fully decentralized system might have individual patients maintain their own data, such as on a personal medical card that includes the entire patient record.[49] Such a system would similarly rely on meaningful standards to ensure transportability and access of patient data by different actors in the health care system.

Any of these systems might potentially work as infrastructure *for* health data, to help enable care. However, a centralized system carries a substantial benefit when considering health data *as* infrastructure for later health innovation.[50] Decentralized data are fragmented along different dimension—not necessarily among different providers and actors in the health system, but between different patients. However, many benefits of health data rely on

---

[47] *See, e.g.*, HOFFMAN, *supra* note 1, at 148–49 (2016) (describing federated databases and their privacy benefits).

[48] *See supra* Section I.B.

[49] *See, e.g.*, Michael Chen & Adrian Gropper, *Patient-Centered EHR Features and Demo*, www.hieofone.org (October 15, 2016) (describing and demoing the concept for an entirely patient-focused individual health record).

[50] *See* OECD, DATA-DRIVEN INNOVATION: BIG DATA FOR GROWTH & WELL-BEING, 177–206 (2015) (applying an infrastructure model for big data generally).

aggregating data from many patients, including precision medicine, quality metrics, and efficiency measures. The risks for health innovation described above include the problems of biases from incomplete data and the risk of innovation being absent altogether. Centralized health data ameliorate these risks by creating comprehensive datasets for future analysis.

Infrastructure goods are typically undersupplied.[51] Infrastructure resources are nonrivalrous inputs into a wide range of output services and goods, with social demand "driven primarily by downstream productive activity,"[52] and substantial spillover benefits.[53] Accordingly, we expect private actors to invest at suboptimal levels in infrastructure spending, suggesting a need for some form of central investment.[54] The federal government is an obvious choice, and indeed the federal government already operates substantial examples of health data infrastructure.[55] These include the multi-site-but-connected Sentinel Project (wherein FDA collects safety information on drugs in use),[56] the Medicare and Medicaid systems, the Veterans Administration,[57] and—specifically focused on forward-looking health research—the Precision Medicine Initiative, aiming to collect comprehensive data on at least one million Americans.[58] An alternate model could rely on public-private partnerships, joining a central government authority with nonprofit actors.[59] However, there is no fundamental requirement that the infrastructure provider be governmental or nonprofit; a for-profit entity can provide public infrastructure given appropriate incentives.[60]

Centralization has complex effects on potential privacy risks. On the one

---

[51] BRETT M. FRISCHMANN, INFRASTRUCTURE: THE SOCIAL VALUE OF SHARED RESOURCES 14–15 (2013) ("society should expect underprovision of [infrastructure] goods.").

[52] *Id.* at xiv.

[53] *Id.* at 5–6.

[54] *Id.* at 14–15.

[55] *See* Eisenberg & Price, *supra* note 2, at __; FRISCHMANN, *supra* note 51, at 14 (noting government provision of goods as a classic solution to infrastructure problems, alongside government subsidies, community provision, and policies to allow private actors to charge supramarginal costs).

[56] See Health Affairs, *Health Affairs Health Policy Brief, The FDA's Sentinel Initiative.* (June 4, 2015), http://healthaffairs.org/healthpolicybriefs/brief_pdfs/healthpolicybrief_139.pdf; Price, *Big Data, Patents, and the Future of Medicine, supra* note 30, at 1441–42 (describing the Sentinel project's data implications); Ryan Abbott, *The Sentinel Initiative as Knowledge Commons*, in GOVERNING MEDICAL COMMONS (BRETT M. FRISCHMANN, MICHAEL J. MADISON, & KATHERINE J. STRANDBURG, EDS.) (forthcoming).

[57] *See* Price, *Big Data, Patents, and the Future of Medicine, supra* note 30, at 1440–41 (describing the Veterans Administration's data).

[58] *Id.* at 1442–43; Francis S. Collins & Harold Varmus, *A New Initiative on Precision Medicine*, 372 N. ENGL. J. MED. 793 (2015).

[59] *See* FRISCHMANN, *supra* note 51, at 14.

[60] Examples include toll-road operators, power companies, and other public utilities. *See id.* Of course, these monopolies raise their own concerns about potential rent-seeking behavior.

hand, centralization creates a broader picture of an individual's health—indeed, that's the point—but that makes it easier to derive more information about an already-identified individual, and also potentially makes it easier to identify a de-identified individual from a larger collection of data.[61]   A centralized system is also a more attractive target for attacks. On the other hand, centralization, or just a coherent infrastructure, allows some privacy-enhancing technologies to be deployed, such as one-way hashing, dataset-docking, or simply scaled security given the possible concentration of resources at a single location.

CONCLUSION

The health system relies on data, but collects and maintains those data in a haphazard, fragmented, and insecure way that creates real risks for patients and for the system as a whole. Given market incentives driving competition among different data systems and health actors, health data seem likely to remain fragmented without broader systemic action. Conceiving of infrastructure both for and of health data suggests that standardized, centralized collection and maintenance of health data may create goods at both the individual and systemic level. If we are to realize the goal of data-informed patient care and data-driven development of future medical technology, an infrastructure for health data provides a substantial step in the right direction.

---

[61] For instance, there may be many people in a particular health system that fit two or three given characteristics; many fewer fit twenty or thirty, and two or three hundred would be much more likely to apply only to a single individual. *Cf.* Orin S. Kerr, *The Mosaic Theory of the Fourth Amendment*, 111 MICH. L. REV. 311 (2012) (noting in the Fourth Amendment context that collections of otherwise non-individualized characteristics can identify an individual).