2017

Risk and Resilience in Health Data Infrastructure

W. Nicholson Price II
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.

* Assistant Professor of Law, University of Michigan Law School. JD, 2011, Columbia University School of Law. PhD (Biological Sciences), 2010, Columbia University Graduate School of Arts and Sciences. For helpful conversations and feedback, I wish to thank Ana Bracic, Rebecca Eisenberg, Brett Frischmann, and the participants in the Silicon Flatirons Digital Broadband Migration Conference. All errors are my own.

INTRODUCTION

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 have the potential 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.[1]

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.[2] 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.[3]

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.[4] 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.[5]

---

1. *See infra* Section I.A.
2. Integrated records also exist when the provision of care is itself integrated rather than fragmented; an integrated care provider of care such as Kaiser Permanente can maintain integrated health records because that one entity provides all aspects of patient care and payment. *See infra* Section I.B.1. Even for integrated providers, however, patient records may be fragmented when patients shift between providers over time.
3. *See infra* Section I.A.
4. *See infra* Part II.
5. *See id.*

Efforts to combat data fragmentation would benefit from looking at health data through an infrastructure lens. I draw on Brett Frischmann's extensive analysis of infrastructure, which he characterizes as largely nonrivalrous resources that derive value principally from their many downstream uses.[6] Most obviously, health data infrastructure would be infrastructure *for* health data—that is, infrastructure on which health data can be stored and transmitted (such as computer systems, shared data standards, and the like). But it should also be 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. In an infrastructure *of* data, the data themselves are a resource to enable productive downstream activity that can improve the health care system.

This essay proceeds in three parts. Part I describes the landscape of health data today, including potential benefits of the collection and analysis of health data and the reasons for fragmentation that limits those benefits. Part II describes the risks that arise from a fragmented health data system. To be clear: this brief essay does not attempt to completely catalog all risks that arise from the use of data in health care; it focuses instead on a subset of particularly salient risks that arise specifically from the problem of fragmentation.[7] Part III sketches the basics of an infrastructure vision for and of health data.

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).[8] New forms of data are proliferating to fill those records, including the reports of traditional medical encounters, high-volume

6. BRETT M. FRISCHMANN, INFRASTRUCTURE: THE SOCIAL VALUE OF SHARED RESOURCES 61 (2012) (describing the three key characteristics of infrastructural resources as nonrivalrous consumption, value derived from input into downstream uses, and the ability to be an input for a wide range of such downstream uses).

7. To take the easiest example, the underlying data may be inaccurate, whether due to errors collecting or entering the data, or may be systematically biased. *See, e.g.*, Sharona Hoffman & Andy Podgurski, *Big Bad Data: Law, Public Health, and Biomedical Databases*, 41 J.L. MED. & ETHICS 56 (2013). If underlying data are inaccurate, joining them into easy-to-use centralized databases will not solve that inaccuracy (though the possibility of cross-checking might ameliorate the problem).

8. 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 (ARRA) of 2009, div. A, tit. XIII, div. b, tit. IV, Pub. L. No. 111-5, 123 Stat. 115 (2009). *See* SHARONA HOFFMAN, ELECTRONIC HEALTH RECORDS & MEDICAL BIG DATA: LAW AND POLICY 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; *Medicare and Medicaid EHR Incentive Program Basics*, CTRS. FOR MEDICARE & MEDICAID SERVS., https://www.cms.gov/regulations-and-guidance/legislation/ehrincentiveprograms/basics.html [https://perma.cc/7DFK-AHW7] (last visited Jan. 12, 2016).

diagnostic tests, such as genetic sequencing and analysis, prescription records, and others.[9]

A. Potential Benefits

These data can create substantial benefits for patients, providers, and for the health system as a whole.[10] Ideally, they should lead to improved care for individual patients as integrated medical records prevent easily avoidable medical errors and allow a broader picture of the patient's overall health.[11] They should enable more efficient care by reducing the costs of coordination, should decrease costs, and should even enable more effective and efficient billing for health treatments. 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.[12] Health data enable the imposition of sanctions or the provision of incentives to try to shape health care in productive ways.[13]

Data can also 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,[14] that some treatments may be more cost-effective at providing the same benefit,[15] that some patients gain more benefit from a particular treatment than others,[16] or that a drug should be moved from prescription-only to over-the-counter status.[17] Recently, the FDA has even gained the statutory authority to use this type of real-world evidence to approve new indications for drugs.[18] More broadly,

9. *See* Rebecca S. Eisenberg & W. Nicholson Price II, *Promoting Health Innovation on the Demand Side*, 4 J.L. & BIOSCIENCES 3 (2017).

10. These multiple uses do not mean that doctors or others collect the data with those purposes in mind; data may be collected just to monitor care, or for the purposes of billing, or for many potential reasons. But once data are collected, they can be used in many different ways.

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

12. *See id.*

13. *See* Medicare Access and CHIP Reauthorization Act (MACRA) of 2015, 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).

14. *See* Eisenberg & Price, *supra* note 9, at 7 (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).

15. *See id.* at 16–18 (describing cost-effectiveness research and the use of observational studies of patient data to perform such research).

16. *See id.* (describing comparative-effectiveness research).

17. *Id.* at 7–10 (describing a petition filed by Blue Cross of California (later Wellpoint) to take certain antihistamines, including Claritin, over-the-counter).

18. *See* 21st Century Cures Act, Pub. L. No. 114-255, § 3022 (2016) (requiring FDA to "establish a program to evaluate the potential use of real world evidence" for the approval

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.[19] 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 data mining, rely on large sets of existing health data.[20]

Overall, health data offer substantial promise for improving health care, in terms of both near-term, patient-specific benefits, and later innovations to improve the health system. Unfortunately, these benefits have been slow to materialize. One cause of this slowness is the fragmentation of health data.[21]

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 (HIPPA), make it hard to link data.

1. Fragmented Care

The key underlying cause of health data fragmentation is that health care is itself frequently fragmented.[22] Patients see different

---

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 NEW ENG. J. MED. 2473 (2015).

19. 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. INFORMATICS ASS'N 413 (2016); Marc L Berger et al., *Opportunities and Challenges in Leveraging Electronic Health Record Data in Oncology*, 12 FUTURE ONCOL. 1261 (2016).

20. *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).

21. 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 9, at 9–10, 27–28 (discussing the problematic incentives for drug manufacturers and health insurers, respectively); David Orentlicher, *Paying Physicians More to Do Less: Financial Incentives to Limit Care*, 30 U. RICH. 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 9, at 23–26. And even once innovative information is generated, getting health care providers to implement the new knowledge can be challenging. *Id.* at 28–32.

22. *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). There are exceptions, including integrated health systems such as Kaiser Permanente, Geisinger, or the federally-run Indian Health Service and Veterans Administration.

doctors at different times, visit different drugstores, change insurers, and in other ways participate in an inherently fragmented health system.[23] 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.[24] 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.[25] 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,[26] 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),[27] Apple (which aims to create a personal digital hub of health information),[28] 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 siloes. This might not be so problematic if communication and data-sharing between the siloes 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

23. *See* Eisenberg & Price, *supra* note 9, at 28–32.

24. *Id.*

25. *Id.* at Section I.D.

26. 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/ [https://perma.cc/F892-3PRU] (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).

27. 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 (Sept. 9, 2016), http://www.nytimes.com/2016/09/11/sports/ncaafootball/wearable-technology-nike-privacy-college-football.html [https://perma.cc/5XEW-7273].

28. *See A Bold New Way to Look at Your Health*, APPLE, http://www.apple.com/ios/health/ [https://perma.cc/QV4M-KVEW] (last visited Oct. 11, 2017) (describing the iOS Health App, which collects phone data and can serve as a repository for personal medical records).

diagnostic tests, for instance, or among others that collect health or health-related data.[29]

In addition to competition among those who generate data, there is competition among the vendors who provide ways of generating and managing data. The electronic health record market is itself fragmented, with hundreds of vendors.[30] This itself could organically lead to fragmentation through lack of 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.[31] 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.[32] The HIPAA Privacy Rule places limits on how personally identifiable health data may be used and disclosed.[33] In general, all uses and disclosures of such information by covered entities—providers, insurers, and health

29. 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 B.U. J. SCI. & TECH. L. 233 (2015) (describing how patents led to Myriad's competitive advantage).

30. *See* OFF. NAT'L COORDINATOR FOR HEALTH INFO. TECH., *Hospital Health IT Developers* (July 2017), https://dashboard.healthit.gov/quickstats/pages/FIG-Vendors-of-EHRs-to-Participating-Hospitals.php [https://perma.cc/P2YC-T9J9]. The top six vendors provide services for 92% of all nonfederal acute-care hospitals. *Id.*

31. *See* OFF. NAT'L COORDINATOR FOR HEALTH INFO. TECH., REPORT TO CONGRESS: REPORT ON HEALTH INFORMATION BLOCKING 11–19 (Apr. 2015), *available at* www.healthit.gov/sites/default/files/reports/info_blocking_040915.pdf [https://perma.cc/8K8M-6E3X] *(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).*

32. Health Insurance Portability and Accountability Act, Pub. L. No. 104-191, 100 Stat. 2548 (1996).

33. HIPAA's principal data restrictions come from the Privacy Rule, codified at 45 C.F.R. § 160 (2016). HIPAA's regulatory structure is complex and need not be discussed in full here; for additional information, *see, e.g.*, U.S. DEP'T. OF HEALTH & HUMAN SERVS., *Summary of the HIPAA Privacy Rule* (May 2003), https://www.hhs.gov/sites/default/files/privacysummary.pdf [https://perma.cc/9SBU-JX9R] (providing HIPAA overview); Eisenberg & Price, *supra* note 9, at 32–35 (discussing the Privacy Rule in the context of research using existing health data).

data clearinghouses[34]—are prohibited unless specifically permitted. To 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 invoking HIPAA as a blanket excuse for refusing to share information, including for uses that are expressly permitted.[35] As Arti Rai describes it, "compliance with the Common Rule [governing research on human subjects] and the HIPAA Privacy Rule imposes a tax on sharing data."[36]

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.[37] As a result, secondary research often involves health information that has been de-identified, which takes it out of HIPAA's ambit.[38] 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.[39] 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.[40] But increasingly, relevant health information is held by entities outside that system, such as 23andMe, Fitbit, Apple, or

34. 45 C.F.R. § 160.103. Uses or disclosures by the business associates of covered entities are governed, though by contract rather than directly under HIPAA's Privacy Rule. 45 C.F.R. § 152(a)(3).

35. 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), https://www.nytimes.com/2015/07/21/health/hipaas-use-as-code-of-silence-often-misinterprets-the-law.html?mcubz=1 [https://perma.cc/L8UG-TRTF].

36. Arti K. Rai, *Risk Regulation and Innovation: The Case of Rights-Encumbered Biomedical Data Silos*, 92 NOTRE DAME L. REV. 1641, 1652 (2017) (noting in addition, "At least for some kinds of data, this tax can be relatively modest.").

37. 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* 21st Century Cures Act, H.R. 6, 114th Cong. § 1124 (2015), *available at* https://www.congress.gov/114/bills/hr6/BILLS-114hr6ih.xml [https://perma.cc/V92W-WGDY]. As passed, the legislation calls instead for the study of such an amendment to the Privacy Rule. 21st Century Cures Act, Pub. L. No. 114-255, § 2063 (2016).

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

39. *See* W. Nicholson Price II, *Big Data, Patents, and the Future of Medicine*, 37 CARDOZO L. REV. 1401, 1413 (2016); *see also* Ryan Abbott, *Big Data and Pharmacovigilance: Using Health Information Exchanges to Revolutionize Drug Safety*, 99 IOWA L. REV. 225, 252–53 (2013) (noting the problem of interoperability for health records).

40. 45 C.F.R. § 160.103.

others.[41] None of these entities, or the data they hold, are directly governed by HIPAA.[42] On the one hand, this might seem to improve the problem of data fragmentation; these entities can gather data unhindered by HIPAA's strictures. On the other hand, fragmentation may increase because different entities, with different forms of health data, are governed by different legal regimes.[43]

Notably, there have also been governmental efforts to encourage 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.[44] And, of course, the push toward electronic health records was itself a federal initiative.[45] Other private systems have been created with the goal of collecting data across providers in order to ensure continuous care and ease the processing of claims; however, these efforts have met with real challenges.[46] Overall, health data in the U.S. health care system remain highly fragmented among different entities, working with different and often mutually incompatible health records systems.

41. "Covered entities" governed by HIPAA include health plans, health information clearinghouses, and health-care providers who transmit certain information electronically. *Id.* Entities like 23andMe, Fitbit, and Apple fit into none of these categories.

42. If these entities are business associates of covered entities, they may be regulated by HIPAA as described in note 34, *supra*.

43. This disparity also raises separate concerns about the fragmentation of *governance* of different health data sources and types. *See* Nicolas Terry, *Regulatory Disruption and Arbitrage in Healthcare Data Protection*, 17 YALE J. HEALTH POL'Y L. & ETHICS (forthcoming 2017).

44. OFF. NAT'L COORDINATOR FOR HEALTH INFO. TECH., CONNECTING HEALTH AND CARE FOR THE NATION: A 10-YEAR VISION TO ACHIEVE AN INTEROPERABLE HEALTH IT INFRASTRUCTURE (2014), *available at* http://www.healthit.gov/sites/default/files/ONC10year InteroperabilityConceptPaper.pdf [https://perma.cc/4FEF-KDXH] [hereinafter ONC, INTEROPERABILITY 10-YEAR VISION]; OFF. NAT'L COORDINATOR FOR HEALTH INFO. TECH., CONNECTING HEALTH AND CARE FOR THE NATION: A SHARED NATIONWIDE INTEROPERABILITY ROADMAP (Draft Version 1.0 April 2015), *available at* http:// www.healthit.gov/sites/default/files/nationwide-interoperability-roadmap-draft-version-1.0.pdf [https://perma.cc/B9J8-KEAX] [hereinafter ONC, INTEROPERABILITY ROADMAP]; *see also* Abbott, *supra* note 39, at 252–53 (noting the efforts of the Office of the National Coordinator in attempting to combat interoperability and fragmentation challenges).

45. American Recovery and Reinvestment Act of 2009, Pub. L. No. 111-5, 123 Stat. 115, 186, 231 (2009).

46. 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* (Aug. 5, 2014), https://www.stjhs.org/ documents/HIE/CalINDEX_news-release.pdf [https://perma.cc/7QG9-NV3L] ("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*, MOD. HEALTHCARE (Mar. 5, 2016), http://www.modernhealthcare.com/article/20160305/ MAGAZINE/303059948 [https://perma.cc/KAZ4-VV9R].

II. RISKS OF THE CURRENT SYSTEM

The risks from fragmentation in the health data system are substantial.[47] These risks come in two main buckets: primary risks, which is to say risks to patients seeking care in the health system, including problems in the quality of patient care and problems for patient privacy; and secondary risks, which is to say risks that arise when health data are repurposed and used to innovate or improve the system. These categories mirror the concept of primary use of data for patient care and secondary use of data for other purposes.[48]

A. Primary Risks

Primary risks from health data fragmentation include both risks to patient health and risks to patient privacy. The risks that arise in patient care mirror the potential benefits of electronic health records (EHRs). If doctors expect patient information to be present in a patient's 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. This works fine if the information is actually present, but decreases the likelihood of catching an error when the information is missing due to fragmentation or otherwise. Such errors can arise from the transitional and fragmented status of such records, where the promise of comprehensive information is provided by EHRs but that promise is not yet realized. This is not to suggest that the previous, paper-based system was impervious to error—far from it—but rather to identify a potential source of error in the current fragmented system.[49]

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.[50] This risk is quite familiar, as it arises from fragmented data whether paper-based or electronic.

---

47. As noted above, other risks exist in the health data system, but are not the focus of this brief essay. *See supra* note 7.

48. *See* OECD, STRENGTHENING HEALTH INFORMATION INFRASTRUCTURE FOR HEALTH CARE QUALITY GOVERNANCE: GOOD PRACTICES, NEW OPPORTUNITIES AND DATA PRIVACY PROTECTION CHALLENGES 22 (2013) [hereinafter OECD, HEALTH INFORMATION INFRASTRUCTURE].

49. *See, e.g.*, INST. OF MED., TO ERR IS HUMAN: BUILDING A SAFER HEALTH SYSTEM (Linda T. Kohn et al. eds., 2000) (noting the problems of medical error in the health system).

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

Lastly, when health data aren't meaningfully collected or linked together, 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 foregone benefits. To take a simple example, suppose that, as part of a research study, a young woman has her genome sequenced;[51] 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,[52] 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.[53] Different actors in the system store information in different ways, 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

51. For the sake of the example, let us assume the lab is certified under the 1967 Clinical Laboratory Improvements Amendments (CLIA), codified as amended at 42 U.S.C. § 263(a), and that the genetic sequencing is thus of high-enough quality to guide clinical care.

52. A substantial literature considers the question of returning results from genetic research, which involves questions of patient preference, the clinical validity and utility of research findings, the nature of the researcher-patient relationship, the question of informed consent, privacy concerns for patients and family about testing for inheritable disease susceptibility, and other challenges. This essay does not address these many issues, instead using the case of genetic testing as an example of a benefit foregone because of data fragmentation. For an introduction to issues in returning results from genetic testing, *see* Susan M. Wolf et al., *The Law of Incidental Findings in Human Subjects Research: Establishing Researchers' Duties*, 36 J. L. 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 GENETICS 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. 22 (2014) (noting different models of returning data and different possibilities for informed consent).

53. *See* Roger A. Ford & W. Nicholson Price II, *Privacy and Accountability in Black-Box Medicine*, 23 MICH. TELECOMM. & TECH. L. REV. 1 (2016) (describing the privacy concerns related to patient health information); Nicolas Terry, *Protecting Patient Privacy in the Age of Big Data*, 81 UMKC L. REV. 385 (2012).

instance, problems with interoperability (and potentially with HIPAA) may be related to the otherwise-baffling persistence of faxed requests for information between different providers—faxes, after all, transmit data as simple images, a lowest-common-denominator format. 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.[54]

B. Secondary Risks

The secondary risks from fragmented data come from efforts to use those data for future innovation.[55] Such efforts include the FDA's Sentinel initiative to monitor drug usage for safety risks,[56] observational studies to drive care (which can potentially be used to approve new drug indications under the 21st Century Cures Act[57]), machine-learning efforts to discover new biological relationships,[58] and implementations of a learning health care system generally.[59] All of these secondary uses of health care data 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.[60] A fundamental data mining principle is "garbage in, garbage out;" when health care fragmentation creates inaccuracies in data later used in innovation—in addition to any inaccuracies that might have existed before—that innovation suffers, and so may future patients.

---

54. 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 mailing, phoning, or faxing to request records from another provider, and provides only contact information to receive information through those means of communication. *See* U. OF MICH. HOSPS. & HEALTH CTRS, REQUEST FOR OUTSIDE RECORDS – PATIENT INFORMATION FROM ANOTHER ORGANIZATION (2013), *available at* http://www.med.umich.edu/him/release-from-other-organizations.pdf [https://perma.cc/A2DE-XUGB].

55. *See generally* Eisenberg & Price, *supra* note 9 (describing potential innovation by health care payers using existing health data).

56. Susan Forrow et al., *The Organizational Structure and Governing Principles of the Food and Drug Administration's Mini-Sentinel Pilot Program*, 21 PHARMACOEPIDEMIOLOGY & DRUG SAFETY 12 (2012).

57. 21st Century Cures Act, Pub. L. No. 114-255, § 3022 (requiring the Secretary of Health and Human Services to "[E]stablish a program to evaluate the potential use of real world evidence . . . to help support the approval of a new indication for a drug . . . .").

58. *See* Price, *supra* note 20.

59. *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).

60. *See, e.g.*, Hoffman & Podgurski, *supra* note 7.

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. The continued fragmentation of health data, where each health system actor is responsible for generating, collecting, and storing the data for its own interactions with patients, suggests that the system needs intervention to avoid ongoing risks. 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 one specific form, but the idea of health data infrastructure, and the risks of fragmented health data, suggest some features of the desired state. This Part briefly describes infrastructural resources, relying on a theoretical framework elaborated by Brett Frischmann.[61] It then identifies how health data fit into an infrastructure model, both infrastructure *for* health data, and infrastructure *of* health data. Finally, it addresses the implications of an infrastructure model for policy interventions regarding health data.

A. Infrastructural Resources

Frischmann has described infrastructure resources as having three principal characteristics. First, an infrastructural resource "may be consumed nonrivalrously for some appreciable range of demand;"[62] that is to say, consumption by one does not decrease the opportunity for consumption by another (within some range beyond which congestion may decrease the resource's usability). This allows "widespread, shared access and productive use of the good," and characterizes it as a pure or impure public good.[63] Second, "[s]ocial demand for the resource is driven primarily by downstream productive activities that require the resource as an input."[64] Roads are not valuable principally because you can drive on them; roads are valuable because you can use them to get places and transport goods.[65] Thus, infrastructure resources are most valuable as means for downstream uses rather than ends of themselves.[66] Third, and finally, "[t]he resource may be used as an input into a wide range of goods and services, which may include private goods, public goods, and social goods."[67] It is not "optimized for a particular user or use;" instead, "[u]sers determine what to [do] with the capabilities that

61. FRISCHMANN, *supra* note 6.
62. *Id.* at 61.
63. *Id.* at 62.
64. *Id.* at 61.
65. *Id.* at 64.
66. *Id.*
67. *Id.* at 61.

infrastructure provide[s]."[68] And because some of the outputs of infrastructure are public goods and social goods, industry is typically undersupplied by markets or, if privately supplied, is focused on an overly narrow set of uses.[69]

B. Infrastructure For and Of Health Data

So what is infrastructure *for* health data? I use this term to describe the set of resources that enable the collection, storage, transmission, and use of health data. And indeed the term "health data infrastructure" or "health information infrastructure" has typically meant just this. Over twenty years ago, Larry Gostin, in an article focused on health information privacy, defined "health information infrastructure" as "the basic, underlying framework of electronic information collection, storage, use, and transmission that supports all of the essential functions of the health care system."[70] Health information infrastructure has been even more broadly defined by a leading national committee as "the values, practices, relationships, laws, standards, systems, applications, and technologies that support all facets of individual health, health care, and population health."[71] Thus, computer systems, data standards, and communication protocols are included within such very broad definitions of health infrastructure.[72] What is *not* included, and is described below, are the data themselves.

This concept of infrastructure for health data largely tracks Frischmann's characterization. These resources are largely

68. *Id.* at 65.

69. *Id.* at 66.

70. Lawrence O. Gostin, *Health Information Privacy*, 80 CORNELL L. REV. 451, 456 (1995) (citing COMM. ON REG'L HEALTH DATA NETWORKS, NAT'L ACAD. SCI., HEALTH DATA IN THE INFO. AGE: USE, DISCLOSURE, & PRIVACY (Molla S. Donaldson & Kathleen N. Lohr, eds. 1994); WORK GROUP ON COMPUTERIZATION OF PATIENT RECORDS, U.S. DEP'T OF HEALTH & HUMAN SERVS., TOWARD A NAT'L HEALTH INFRASTRUCTURE (1993); KAREN A. DUNCAN, HEALTH INFO. & HEALTH REFORM: UNDERSTANDING THE NEED FOR A NAT'L HEALTH INFO. SYSTEM (1994)). A decade earlier, a National Library of Medicine planning panel proposed the idea of U.S. health infrastructure based on "a national computer network for use by the entire biomedical community, both clinical and research professionals." NAT'L LIBRARY OF MED. PLANNING PANEL NO. 4: LONG RANGE PLAN ON MED. INFORMATICS (1986) (quoted in Don E. Detmer, *Building the National Health Information Infrastructure for Personal Health, Health Care Services, Public Health, and Research*, 3 BMC MED. INFORMATICS & DECISION MAKING 3 (2003). *See also* JASON, *A Robust Infrastructure Health Data Infrastructure*, AGENCY FOR HEALTHCARE RESEARCH & QUALITY (Nov. 2013), https://www.healthit.gov/sites/default/files/ptp13-700hhs_white.pdf [https://perma.cc/WP8C-MRT8] (noting the history of calls for health data infrastructure).

71. NAT'L COMM. ON VITAL & HEALTH STATS., INFO. FOR HEALTH: THE STRATEGY FOR BUILDING THE NAT'L HEALTH INFO. INFRASTRUCTURE 11 (2001); *see also id.* at 16 (elaborating on this definition and describing the goals of the National Health Information Infrastructure); *see also* JASON, *supra* note 70, at 1 (describing "a combination of electronic health records (EHRs) and improved exchange of health information" as "health data infrastructure").

72. *See also* OECD, HEALTH INFO. INFRASTRUCTURE, *supra* note 48, at 13 (not defining health data infrastructure but focusing on electronic health records and ways to link and connect them).

nonrivalrous—they can be used by many simultaneously without diminishing the benefits for others, and in fact increased use increases their value as networks and as standards to promote interoperability. In addition, their value lies not in their own use—standards to encode health data or computer systems to store them are not valuable on their own, but because of the primary and secondary uses they enable for health data.[73] And those uses are many, including "clinical and prevention services, quality assurance, financial reimbursement, monitoring of fraud and abuse, research, and public health services."[74]

As I have described above, however, we can also conceive of an infrastructure *of* health data—that is, the view that the data themselves are an infrastructural resource. But in fact, health data also fit the characteristics of an infrastructural resource as laid out by Frischmann.[75] Health data are largely nonrivalrous, like other information goods—my use of a set of treatment outcomes to conduct innovative research does not interfere with your use of that same set of treatment outcomes to measure the quality and efficiency of the health care system.[76] Health data are principally valuable for their downstream uses—it may be interesting to know one's cholesterol levels, but those data are truly important for what users, whether patients, doctors, or researchers, can do with them.[77] And finally, those downstream uses of health data are highly variable—doctors can use health data to direct treatment for an individual patient, researchers can use health data to develop new drugs or treatments, and administrators can use health data to measure system quality and develop incentives to improve that quality, among many other possibilities.[78]

---

73. One could argue whether primary use of patient data by providers really qualifies as "downstream" use of the health infrastructure. If not, then this set of resources merely has a partially infrastructural quality, but the same arguments still apply, though possibly with lesser force. As noted above, these resources are generally described as infrastructural by scholars and policymakers.

74. Gostin, *supra* note 70, at 456.

75. For an application of Frischmann's infrastructure model to big data more generally, *see* OECD, DATA-DRIVEN INNOVATION: BIG DATA FOR GROWTH & WELL-BEING 177–206 (2015) (hereinafter OECD, DATA-DRIVEN INNOVATION).

76. *See id.* at 179–80 (describing big data as nonrivalrous).

77. *See id.* at 180–81 (describing big data as a capital good). As above, one could argue that health data uses at the point of care—that is, primary uses—are not, in fact, "downstream," and that such value could be internalized at the point of care. Accepting this argument would mean that health data only have some infrastructural characteristics, instead of being fully infrastructural resources. The arguments about provision, governance, and the like described below still hold, though their magnitude may be decreased. *Cf. id.* at 63 (even though one can derive some consumption value from roads from driving for fun, they create most of their value through downstream uses and are thus infrastructural).

78. *See id.* at 181–83 (describing big data as a general-purpose input); Eisenberg & Price, *supra* note 9, at 14–23 (describing several of the uses to which health data can be put by one type of user, health insurers).

C. *Implications of an Infrastructure Model*

What does an infrastructure model both *for* and *of* health data imply? I draw three primary implications from applying an infrastructure model: likely government involvement in provision, relatively open access, and a preference for centralization.

1. Government Involvement

Infrastructural resources are typically undersupplied by the private sector.[79] Because they are inputs into a broad set of uses that include public and social goods, with typically substantial spillovers, it is difficult for private actors to capture the full value of investing in infrastructure.[80] Accordingly, we expect private actors to invest at suboptimal levels in infrastructure spending, suggesting a need for some form of governmental investment or subsidy.[81] The federal government is an obvious choice as the largest payer for health care and the entity with the possibility to break down state-by-state siloes of data, and indeed the federal government already operates substantial examples of health data infrastructure.[82] With respect to infrastructure *for* health data, the federal government has long been involved in developing that infrastructure, including several substantial panels and reports.[83] Most recently, the federal government committed billions of dollars in incentives in the HITECH Act for the adoption of electronic health records, and created corresponding penalties for failure to adopt them.[84] Nevertheless, the government has also taken a lighter touch in some areas of infrastructure for health data—it has forcefully stated the case for interoperability, but declined to mandate the standards that would make such interoperability more straightforward.[85] As a result, electronic health record formats are still frequently

79. FRISCHMANN, *supra* note 6, at 14–15 ("society should expect underprovision of [infrastructure] goods.").

80. *Id.* In addition, because demand-side users of infrastructure are often unable to capture the full value of public goods or social goods that they produce, demand for infrastructure resources, as indicated purely by competitive markets, may also decrease the supply of infrastructure resources. *Id.* at 72–78.

81. *Id.* at 14–15.

82. *See* Eisenberg & Price, *supra* note 9, at 40; FRISCHMANN, *supra* note 6, 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). The federal government is not the only choice; states have the primary role in regulating health, and might be another option. But state-by-state regimes risk replicating fragmentation on a state level, and ERISA limits state abilities to regulate the activities of some health actors such as many employer-funded health insurance plans. *See* Gobeille v. Liberty Mut. Ins. Co., 136 S. Ct. 936 (2016) (holding a Vermont law requiring insurers and providers to report claims data to a state-run database was preempted by ERISA).

83. For a few of the more high-profile reports, *see supra* notes 70–72.

84. *See supra* note 8.

85. *See* ONC, INTEROPERABILITY ROADMAP, *supra* note 44; ONC, 10-YEAR VISION, *supra* note 44.

incompatible, sometimes deliberately so, hampering the project of infrastructure for health data.[86]

The federal government has also been involved in several initiatives aimed—explicitly or not—at developing an infrastructure *of* health data. These include the multi-site-but-connected Sentinel Project (wherein FDA collects safety information on drugs in use),[87] the Medicare and Medicaid systems, the Veterans Administration,[88] and—specifically focused on forward-looking health research—the Precision Medicine Initiative, aiming to collect comprehensive data on at least one million Americans.[89]

An alternate model to direct federal investment could rely on public-private partnerships, joining a central government authority with nonprofit actors.[90] There is no fundamental requirement that the infrastructure provider be governmental or nonprofit; a for-profit entity can provide public infrastructure given appropriate incentives.[91] But relying on private actors, even with incentives, can reduce spillover benefits, as described in the next section. Overall, an infrastructure model *for* and *of* health data suggests at least some role for government involvement.

2. Openness of Access

A second implication of an infrastructure model is that the infrastructural resources might usefully be governed under a model that permits relatively open access to the resource. As Frischmann notes, infrastructure users frequently produce public goods and social goods (in this case including downstream technical innovation or information about system efficiency).[92] Because they cannot capture the social value of these goods, they are less willing to pay for access to the infrastructural resource than is socially desirable.[93] This creates demand problems, so that using a competitive market to

86. *See supra* note 31 and accompanying text.

87. *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 [https://perma.cc/E9UM-M2GB]; Price, *supra* note 39, 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).

88. *See* Price, *supra* note 39, at 1440–41 (describing the Veterans Administration's data).

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

90. *See* FRISCHMANN, *supra* note 6, at 14.

91. 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.

92. *Id.* at 68–69.

93. *Id.* at 69. As Frischmann notes, although the classical approach to decreasing this public goods problem is to subsidize the production of that public good, subsidizing the production of the infrastructural input may also help. *Id.* at 71; *see also supra* Section III.C.1 (describing such government subsidization of infrastructure).

regulate access to the infrastructural resource is likely to result in an access regime that is tighter than socially optimal.[94] Frischmann suggests that a commons model, wherein access to the resource is available on nondiscriminatory terms to members of the relevant community—which may be the public at large—may be appropriate for many infrastructure resources.[95] For publicly managed infrastructure, as I have suggested health data infrastructure is likely to be and likely should be, Frischmann argues that commons management is particularly appropriate because it creates a "spillover-rich environment" by: (1) allowing users to decide how to use the resource rather than picking beforehand which uses will be prioritized, and (2) sustaining the generic, rather than specialized, nature of the resource, which supports a broad range of potential uses.[96] For health data infrastructure, these reasons suggest that making health data, and the infrastructural resources underlying those data, broadly available for a wide range of uses is likely to produce the greatest public benefit, whether to individual patient care, systemic evaluation, innovation to produce new medical products or new medical knowledge, or independent evaluations of those innovations.[97]

To take a simplified example, imagine a health system implementing a database for secondary use. One version it could implement includes many health variables; another includes only variables relevant to detecting insurance fraud. Potential users include wealthy insurers and poor researchers. Insurers would likely pay the same for either database; if the health system gauged what system it should implement by market demand, and charged for access, it would likely go with the simpler database. But that database would be less socially valuable than a broader database that could enable research use—and other potential downstream uses. Keeping the resource generic, and keeping it broadly available, increases the possibilities for social benefit. But that raises the question of who will pay, which brings us back to the government subsidy point made above.

Broad access to infrastructure *for* health data seems to raise relatively few red flags, but broad access to infrastructure *of* health data—that is, to the data themselves—raises the possibility of substantial privacy concerns. As Roger Ford and I have previously noted, the large amounts of health data useful for developing

94. FRISCHMANN, *supra* note 6, at 71.

95. *Id.* at 92–93. Frischmann notes that whether any particular infrastructure resource should be managed as commons "remains[s an] incredibly difficult question[] that must be answered contextually." *Id.* at 93.

96. *Id.* at 94.

97. *See, e.g.*, Price, *supra* note 39 (describing the possibility of using big health data to develop sophisticated algorithms for use in health care); Ford & Price, *supra* note 53 (describing the possibility of independent evaluation of such algorithms).

machine-learning-based algorithms in health care include sensitive data for about which many worry.[98] Addressing these concerns—or at least considering them carefully and weighing their magnitude—is an important aspect of health data infrastructure.[99] At the very least, addressing privacy concerns may increase the likelihood of voluntary patient buy-in to the idea of broadly sharing health data.[100]

3. Centralization

Finally, an infrastructure model raises the issue of centralization, especially for an infrastructure *of* health data. At one end of the spectrum, 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.[101] This model is closest to the current system—but that closeness demonstrates potential problems, since even with federal initiatives to drive interoperability, fragmentation persists.[102] At the other end of the spectrum, 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.[103] 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 likely carries a substantial benefit when considering health data *as* infrastructure for later health innovation.[104] Decentralized data are

98. Ford & Price, *supra* note 53; *but see* Carl F. Schneider, *A Comment on Privacy and Accountability in Black-Box Medicine*, MICH. TELECOMM. & TECH. L. REV. (forthcoming 2017) (arguing these privacy concerns may be overblown).

99. *See* Gostin, *supra* note 70, at 485–89 (noting the need to consider health information privacy within the context of infrastructure for health data); JASON, *supra* note 70, at 31–34 (focusing on privacy in infrastructure for health data).

100. *See* JASON, *supra* note 70, at 31 (discussing the need for patient trust). Of course, some patients see little need for health data privacy, and willingly share their information publicly. *See, e.g.*, THE PERSONAL GENOME PROJECT, http://personalgenomes.org [https://perma.cc/MZ9W-SJY5] (last visited June 10, 2017) (creating a database for individuals to publicly share their genomic and health data).

101. *See, e.g.*, HOFFMAN, *supra* note 8, at 148–49 (describing federated databases and their privacy benefits). The Sentinel system follows this model. *Id.*

102. *See supra* Section I.B.

103. *See*, *e.g.*, Michael Chen & Adrian Gropper, *Patient-Centered EHR Features and Demo*, (Oct. 15, 2016), http://www.hieofone.org [https://perma.cc/PV2Y-9X8W] (describing and demoing the concept for an entirely patient-focused individual health record); Gostin, *supra* note 70, at 461–63 (describing, in 1995, the potential storage of health data on electronic health record cards).

104. I am certainly not the first to argue for a centralized health data system. *See*, *e.g.*, Gostin, *supra* note 70, at 463–70 (discussing several limited health databases, federal and otherwise, and discussing the possibility of a national data collection initiative). The claim

fragmented along different dimensions—not necessarily among different providers and actors in the health system, but between different patients. However, many benefits of health data rely on 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.

Centralization standing alone also raises concerns about limited competition (if there is only one resource, there is no competition by definition) and about limited access (if actors are shut out of the single resource, where else can they turn?). These concerns are lessened by the considerations suggested before of government involvement and commons management. If a centralized health database is government run or subsidized, competition is of limited use—and, as noted, competition is often insufficient to adequately drive the creation of infrastructural resources.[105] Similarly, effective commons management largely forecloses the problem of limited access—again, reflecting the reality that market-driven demand for infrastructural resources is often insufficient to reflect the social benefits that come from their broad use.[106]

Centralization has complex effects on potential privacy risks. On the one 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.[107] A centralized system is also a more attractive target for attacks and hacking attempts. On the other hand, centralization, or just a coherent infrastructure, allows some privacy-enhancing technologies to be deployed, such as one-way hashing.[108] From a security standpoint, a centralized resource is a more attractive target, but can also be the subject of substantially

of benefits from a federal system is ultimately an empirical claim, and would need to be studied further before making policy decisions.

105. *See supra* Section III.C.1.

106. *See supra* Section III.C.2.

107. 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).

108. *See, e.g.*, Ioana Danciu et al., *Secondary Use of Clinical Data: The Vanderbilt Approach*, 52 J. BIOMED. INFORMATICS 28 (2014) (discussing privacy-protecting practices to store and collect data at Vanderbilt); Abel N. Kho et al., *Design and Implementation of a Privacy Preserving Electronic Health Record Linkage Tool in Chicago*, 22 J. AM. MED. INFORMATICS ASS'N. 1072–80 (2015) (discussing similar practices in Chicago); Ford & Price, *supra* note 53, at 36–37 (discussing protecting health data privacy through technological measures generally).

more security given the possible concentration of resources at a single location. Overall, the case for centralization is not ironclad, but the increased benefits make it strongly worth considering.

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, subsidized and managed as a commons, may create substantial 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 both for and of health data provides a step in the right direction.