DATA-RICH AND KNOWLEDGE-POOR: HOW PRIVACY LAW PRIVATIZED MEDICAL DATA AND WHAT TO DO ABOUT IT

*Louis Enriquez-Sarano**

The Health Information Technology for Economic and Clinical Health Act (HITECH) successfully encouraged widespread adoption of electronic health records (EHR). Their suitability for "big data" analysis make EHR data immensely valuable for secondary research, which could help scientists develop new drugs, medical devices, and public-health knowledge. Thus far, EHR data have not been widely available to academic medical scientists in quantities sufficient to support big data analysis. Instead, the data are aggregated, analyzed, and sold by insurance companies, EHR vendors, and other medical informatics firms. This Note argues that the advent of the EHR data market is a direct result of HITECH's interaction with the Health Insurance Portability and Accountability Act of 1996 (HIPAA) (together, the "Privacy Regime"). The Privacy Regime (1) establishes the necessary preconditions for the EHR data market; (2) funnels EHR data towards a few large firms; and (3) prevents others, including academic scientists, from acquiring data in similarly large quantities.

The Privacy Regime has radically changed medical research regulation. Traditional clinical trials and retrospective studies are governed by the familiar safeguards of medical ethics including IRB review, peer review, and publication. But under the Privacy Regime, private-sector EHR-based studies are not subject to any ethical review. This result subverts the fundamental principles of medical ethics and inhibits socially valuable public-sector research. This Note proposes reforming the Privacy Regime to subject all medical research to ethical review and to incentivize private firms to share EHR data with academic researchers.

* J.D. Candidate 2021, Columbia Law School. The author would like to thank Professor Thomas Merrill for his wise counsel throughout the writing process, and his family, Dr. Veronique Roger, Dr. Maurice Enriquez-Sarano, and Charles-David Enriquez-Sarano, for their enduring support and professional insights. The author also thanks Charles Nathan, Mollie Weiss, Janna Yu, and the staff of the *Columbia Law Review* for their invaluable editorial assistance.

I swear by Apollo the physician . . . and all the gods and goddesses as my witnesses, that, according to my ability and judgement, I will keep this Oath and this contract: . . . Whatever I see or hear in the lives of my patients, whether in connection with my professional practice or not, which ought not to be spoken of outside, I will keep secret, as considering all such things to be private.[1]

INTRODUCTION

Thirty years ago, whenever a cancer patient permitted their doctor to physically examine them or peer inside their body with an x-ray, the resulting images, measurements, and notes would remain in the patient's paper

1. The Hippocratic Oath, NIH, https://www.nlm.nih.gov/hmd/greek/greek_oath.html [https://perma.cc/MP7X-R5LU] (last visited Nov. 3, 2019).

medical record.[2] During a similar examination today, doctors and nurses record this information in the patient's electronic health record (EHR),[3] generating data that can immediately be analyzed and sold by companies unfamiliar to most patients.[4] In contrast to paper records, EHRs are readily accessible not only to care providers, but also to medical insurance companies, EHR vendors, and other firms.[5] Patients may not elect to use paper records instead.[6] Over the past decade, the sale of privately conducted research using "de-identified" EHR data has become a multibillion-dollar industry, operating without any ethical or scientific oversight.[7] Meanwhile, efforts at harnessing this data for academic research have floundered, despite its lifesaving potential.[8] This failure has only grown more troubling during recent months given the possibility of using EHR data to help scientists understand and contain viral outbreaks.[9]

This Note proceeds in three parts. Part I summarizes EHRs' key features, the EHR data market, and the core provisions of the Health Insurance Portability and Accountability Act of 1996 (HIPAA)[10] and the Health Information Technology for Economic and Clinical Health Act

---

2. See infra section I.A.1 (noting that paper medical records were the most common form of record until 2009).

3. EHRs are "digital version[s] of . . . patient[s'] paper chart[s]." What Is an Electronic Health Record (EHR)?, Off. of the Nat'l Coordinator for Health Info. Tech., https://www.healthit.gov/faq/what-electronic-health-record-ehr [https://perma.cc/8S78-HCRD] (last updated Sept. 10, 2019). A patient's physicians and nurses reference their EHR as they would paper records throughout the provision of medical care. See id.; see also infra section I.A.1 (describing the EHR in detail).

4. See infra section I.A.2 (describing the private market for EHR data and EHR-based research).

5. See infra section I.A.2. EHR vendors develop and sell EHR platforms to healthcare providers. How Do I Select a Vendor?, Off. of the Nat'l Coordinator for Health Info. Tech., https://www.healthit.gov/faq/how-do-i-select-vendor [https://perma.cc/Q9YH-54FP] (last updated Oct. 17, 2019). EHR vendors and insurance companies can access EHR data thanks to the HIPAA Privacy Rule. See infra notes 56–62 and accompanying text.

6. See infra notes 63–64, 78–81 and accompanying text (noting the approximately ninety percent EHR penetration rate and the absence of regulations governing the use of de-identified EHR data). In contrast, patients may give or withhold their consent for many uses of their identifiable EHR data. See, e.g., 45 C.F.R. §§ 164.508–510, 164.522 (2019).

7. See infra section I.A.2. De-identified data are EHR data that have been stripped of identifying information such as patient names, addresses, and social security numbers. See infra note 63 and accompanying text. De-identified EHR data are especially valuable because they permit private companies to analyze patient-derived biomedical information without any of the expensive ethical strictures applicable to traditional clinical studies and research using identifiable patient data. See infra notes 112–121 and accompanying text.

8. See infra notes 185, 188 and accompanying text.

9. See Hongzhang Zheng, William H. Woodall, Abigail L. Carlson & Sylvain DeLisle, Can Long-Term Historical Data from Electronic Medical Records Improve Surveillance for Epidemics of Acute Respiratory Infections? A Systematic Evaluation, PLoS One, Jan. 2018, at 2, 10, 11 (discussing how EHR databases could help governments identify and respond more swiftly to novel viruses, including coronaviruses).

10. Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (1996) (codified as amended in scattered titles of the U.S.C.).

(HITECH)[11]—laws referred to collectively in this Note as the "Privacy Regime."[12] Part II then argues that the Privacy Regime created the EHR data market and is harmful because it not only fails to make the vast majority of EHR data available for academic research, but also permits for-profit research without any scientific or ethical scrutiny. Finally, Part III proposes requiring BigMedTech firms[13] to regularly report on their data collection practices, submit their research to independent ethical review, and make data available for academic research. In essence, Congress should offer these firms a bargain: In exchange for continued permission to monetize de-identified EHR data, they must play by the rules of medical research ethics and share the data's benefits with society at large.

This Note contributes to the medical data regulation literature by causally linking private EHR-based research to an existing regulatory regime. Hopefully, noting that the Privacy Regime creates two tracks for EHR-based research (one with and another without ethical guidelines) will add urgency to widespread calls for regulatory reform.[14] Finally, this analysis should serve as a warning to regulators and legislatures around the country contemplating increased privacy protections. Individual privacy must be defended, but it should not come at the cost of a wholesale transfer of valuable and powerful information to an unaccountable private sector.

I. EHR DATA MARKETS, HITECH, AND HIPAA

Before exploring how the Privacy Regime's legal mechanics shape the EHR data market, some factual and legal background is required. Section I.A briefly summarizes the history of medical records, explains their research value in both the private and academic settings, and describes the current state of the EHR data market. Section I.B then summarizes HIPAA, HITECH, and their constituent regulations.

A. *Medical Records and the EHR Data Market*

Section I.A.1 describes EHRs and explains their value to both academia and BigMedTech. Section I.A.2 then describes the state of the EHR data market and its main players.

---

11. Health Information Technology for Economic and Clinical Health Act, Pub. L. No. 111-5, 123 Stat. 226 (2009) (codified in scattered sections of 42 U.S.C.).

12. This Note refers to this system as the "Privacy Regime" because these laws regulate medical data by protecting patient privacy instead of by regulating data flows for secondary uses—that is, how data are used after their use for the provision of care. See infra sections I.B–II.A.

13. For the sake of convenience, this Note refers collectively to insurance companies, EHR vendors, and other firms engaged in the collection and sale of EHR data as "BigMedTech." See infra section I.A.2 for a more detailed account of the EHR data market's key participants.

14. See infra note 190 and accompanying text.

1. *Electronic Health Records and the Promise of Big Data.* — EHRs were not invented in 2009, but in that year, HITECH made them nearly ubiquitous.[15] Untethered from the spatial limits of paper charts, EHRs promised greater quantities of high-quality data, which could be used to better inform clinical decisions. An EHR includes a patient's administrative and billing data, insurance information, demographic information (age, gender, ethnicity, address), vital signs (blood pressure, pulse rate, respiratory rate, body temperature),[16] medical history, family history, medication history, records of hospitalization, surgical and other procedural histories, diagnoses, immunization dates, allergies, radiological images, laboratory test results, and free-text physician notes.[17] EHRs are of tremendous benefit to patients because they enable more efficient collection and communication of patient health information relative to paper records.[18] For instance, if a New Yorker is hit by a car in Denver, it is now much more likely that the physicians treating them will have access to potentially lifesaving information about their preexisting medical conditions or drug allergies. This leads to "improvements in the quality of care[] [and] a reduction in medical errors."[19]

---

15. See infra section I.B.2.

16. Vital Signs (Body Temperature, Pulse Rate, Respiration Rate, Blood Pressure), Johns Hopkins Med., https://www.hopkinsmedicine.org/health/conditions-and-diseases/vital-signs-body-temperature-pulse-rate-respiration-rate-blood-pressure [https://perma.cc/836J-WU3M] (last visited Oct. 19, 2019).

17. See Kelly Devers, Bradford Gray, Christal Ramos, Arnav Shah, Fredric Blavin & Timothy Waidmann, Urban Inst., The Feasibility of Using Electronic Health Data for Research on Small Populations 57 (2013), https://www.urban.org/sites/default/files/publication/22266/413010-the-feasibility-of-using-electronic-health-data-for-research-on-small-populations.pdf [https://perma.cc/6B44-D68S]; Peter B. Jensen, Lars J. Jensen & Søren Brunak, Mining Electronic Health Records: Towards Better Research Applications and Clinical Care, 13 Nature Revs. Genetics 395, 397–99 (2012); Margaret Rouse, Electronic Health Record (EHR), SearchHealthIT, https://searchhealthit.techtarget.com/definition/electronic-health-record-EHR [https://perma.cc/6295-JJQL] (last visited Oct. 19, 2019); What Information Does an Electronic Health Record (EHR) Contain?, Off. of the Nat'l Coordinator for Health Info. Tech., https://www.healthit.gov/faq/what-information-does-electronic-health-record-ehr-contain [https://perma.cc/962L-K3RN] (last visited Oct. 19, 2019).

18. See What Are the Advantages of Electronic Health Records?, Off. of the Nat'l Coordinator for Health Info. Tech., https://www.healthit.gov/faq/what-are-advantages-electronic-health-records [https://perma.cc/8GUU-9ZVF] (last visited Jan. 18, 2020).

19. See Nir Menachemi & Taleah H. Collum, Benefits and Drawbacks of Electronic Health Record Systems, 4 Risk Mgmt. & Healthcare Pol'y 47, 48 (2011); accord Improved Patient Care Using EHRs, Off. of the Nat'l Coordinator for Health Info. Tech., https://www.healthit.gov/topic/health-it-basics/improved-patient-care-using-ehrs [https://perma.cc/K6X6-BEVG] (last visited May 24, 2020) (explaining that EHRs "can improve health care quality" in part because they allow the health information of patients to be "available in one place, when and where it is needed").

EHRs' advantages over paper records in the clinical-care setting apply with even greater force to their "secondary" research uses.[20] Comprehensive EHR systems contain both cross-sectional data and valuable longitudinal data, which can be analyzed more rapidly and at much greater scale than data held in paper records.[21] The data's massive scale makes them immensely valuable once they are aggregated, formatted, and combined with the power of contemporary machine-learning techniques, colloquially known as "big data."[22] Big data represents a leap in analytical possibilities, similar to the advent of random sampling decades ago; scientific investigators can now study not "just a small subset of the relevant data, but close to if not all of it" to test thousands of hypotheses at once.[23] Academic and private-sector scientists hope that this research will lead to the development of novel treatments, drugs,[24] and devices.[25] The possibility of tracking outbreaks of novel coronaviruses, such as COVID-19, presents a

20. See Meredith Nahm Zozus, Rachel Richesson, W. Ed Hammond & Gregory E. Simon, Acquiring and Using Electronic Health Record Data, Duke Univ.: Rethinking Clinical Trials® (Nov. 3, 2015), https://sites.duke.edu/rethinkingclinicaltrials/acquiring-and-using-electronic-health-record-data [https://perma.cc/Z83P-LJHM] ("The use of EHR data collected during the course of clinical care for research purposes is often referred to as a *secondary use* of healthcare data—that is, the data were first collected as part of routine patient care and will be secondarily used for research.").

21. Cross-sectional data are data collected on multiple subjects (individual patients) at a particular point in time, whereas longitudinal data follow the same group of subjects over multiple points in time, making them invaluable to the study of disease progressions and outcomes. See Edward Joseph Caruana, Marius Roman, Jules Hernández-Sánchez & Piergiorgio Solli, Longitudinal Studies, 7 J. Thoracic Disease E537, E537 (2015) ("[Longitudinal data are] particularly useful for evaluating the relationship between risk factors and the development of disease, and the outcomes of treatments over different lengths of time."); see also Jensen et al., supra note 17, at 395 ("[I]ntegrated patient data constitute a computable collection of fine-grained longitudinal phenolypic profiles, facilitating cohort-wide investigations and knowledge discovery on an unprecedented scale.").

22. See, e.g., R. Scott Evans, Electronic Health Records: Then, Now, and in the Future, 2016 Y.B. Med. Informatics S48, S52; Jensen et al., supra note 17, at 399–401 (discussing the varied applications of machine learning to EHR databases including population-level research and integration into clinical prediction and decisionmaking support systems); see also Viktor Mayer-Schönberger, Big Data for Cardiology: Novel Discovery?, 37 Eur. Heart J. 996, 996 (2015) ("Big Data is reshaping the scientific method, and by extension scientific fields, especially those that are data rich such as cardiology.").

23. See Mayer-Schönberger, supra note 22, at 996–97.

24. See Lixia Yao, Yiye Zhang, Yong Li, Philippe Sanseau & Pankaj Agarwal, Electronic Health Records: Implications for Drug Discovery, 16 Drug Discovery Today 594, 594–97 (2011) (identifying three pathways for such discoveries to be made: first, "[u]sing EHRs to identify novel disease relationships"; second, "[a]pplying EHRs for drug usage re-evaluation"; and third, "[f]inding genotype-phenotype associations from EHRs").

25. See Hanae Armitage, Medical Device Safety in the Real World: Tapping EHR Data, Stan. Med. News Ctr. (Oct. 7, 2019), https://med.stanford.edu/news/all-news/2019/10/medical-device-safety-in-the-real-world–tapping-ehr-data.html [https://perma.cc/QWL2-MNSR] ("Researchers used artificial intelligence and de-identified data from electronic health records to identify the safest types of hip implants.").

prime example of how big data EHR analyses could save lives in ways that traditional clinical studies could not.[26]

2. *The EHR Data Market.* — By 2017, large firms were already generating vast profits by selling research conducted on de-identified EHR data (or access to the data) to pharmaceutical companies and others. Social scientist Adam Tanner documents the EHR data market in his book, *Our Bodies, Our Data: How Companies Make Billions Selling Our Medical Records*.[27] Because the industry thrives on secrecy, Tanner's book cannot reveal the EHR market's precise value,[28] but it shows that nearly every firm with access to patient data is joining the gold rush.

The players in the EHR data market include every type of healthcare institution. The major insurers have created or acquired subsidiaries to monetize their data portfolios.[29] Large healthcare providers are getting into the business as well: Kaiser Permanente, Mayo Clinic, and Geisinger Health have all entered the market on their own or partnered with other firms.[30] IQVIA, the largest independent (in that it does not provide direct healthcare services) health-data broker, and its subsidiaries capture "over 33 million records for unique, de-identified patients" from which they generated $30 million in 2015.[31] In 2016, IBM entered the EHR data business by acquiring three firms with a combined total of 310 million patient records.[32]

EHR vendors will likely emerge as the dominant players in the EHR data market thanks to network effects, which generate data at the scale and scope necessary to dominate data markets. To briefly summarize the phenomenon: A product or service that benefits from network effects becomes

---

26. See Zheng et al., supra note 9, at 3–10 (finding that various virus detection algorithms could accurately predict viral respiratory-illness outbreaks and potentially accelerate government response times).

27. Adam Tanner, Our Bodies, Our Data: How Companies Make Billions Selling Our Medical Records 3–4 (2017) (describing the EHR data market, the concept of de-identification, and the investigative work undertaken to write the book).

28. See id. at 139–40 ("[M]ost companies make it difficult or nearly impossible to find out what they do with data from their patient record systems . . . .").

29. UnitedHealth has Optum and 216 million patient records. Data: Establish the Foundation for Better Outcomes, Optum, https://www.optum.com/solutions/data-analytics/data.html [https://perma.cc/K2CW-J7PW] (last visited Dec. 10, 2019). Anthem has HealthCore and forty-eight million records. Jessica Kent, Anthem's HealthCore Offers Big Data Access for Healthcare Analytics, Health IT Analytics (May 29, 2018), https://healthitanalytics.com/news/anthems-healthcore-offers-big-data-access-for-healthcare-analytics [https://perma.cc/LK3Y-STF9]. Blue Cross Blue Shield has Blue Health Intelligence and 125 million records. See Tanner, supra note 27, at 73.

30. See Tanner, supra note 27, at 73.

31. Id. at 139 (internal quotation marks omitted) (quoting Electronic Health Records, IMS Health, https://web.archive.org/web/20151030162107/https://www.imshealth.com/vgn-ext-templating/v/index.jsp?vgnextoid=e1f6e590cb4dc310VgnVCM100000a48d2ca2RCRD&vgnextfmt=default%3E [https://perma.cc/7W5X-EGLG] (last visited Oct. 30, 2015)).

32. See id. at 73–74.

more valuable as more people use it, thus attracting even more users.[33] This creates a powerful feedback loop, which Facebook, Google, telecommunications giants, and many other companies have used to generate massive wealth, both in money and data.[34] Scholars have long recognized that EHR vendors similarly benefit from network effects—as more physicians adopt a certain EHR, even more follow suit.[35] This factor has almost certainly contributed to the concentration of the EHR vendor market: In 2018, Cerner and Epic provided EHR systems for over fifty percent of hospitals in the United States, and for over seventy-five percent of hospitals with more than 500 beds.[36] The network effects feedback loop—by contributing to the dominance of two EHR vendors—leads inexorably to vast concentrations of data. Epic long stood out for refusing to enter the EHR data market. But the company recently announced its project Cosmos, which will mine Epic's 230 million patient records.[37] Cerner, which has about 150 million patient records, already permits customers to perform analyses on "data enclaves," without seeing the data directly.[38] Practice

33. See Michael L. Katz & Carl Shapiro, Systems Competition and Network Effects, 8 J. Econ. Persps. 93, 94 (1994); The World's Most Valuable Resource Is No Longer Oil, but Data, Economist (May 6, 2017), https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data [https://perma.cc/U2W4-6H9P].

34. See Justus Haucap & Ulrich Heimeshoff, Google, Facebook, Amazon, eBay: Is the Internet Driving Competition or Market Monopolization?, 11 Int'l Econ. & Econ. Pol'y 49, 49–54 (2014).

35. See, e.g., Douglas J. Ayers, Nir Menachemi, Zo Ramamonjiarivelo, Michael Matthews & Robert G. Brooks, Adoption of Electronic Medical Records: The Role of Network Effects, 18 J. Prod. & Brand Mgmt. 127, 130–32 (2009). EHR systems are valuable as clinical tools because they can reduce costs as patients move between physicians by eliminating duplicated medical history and data entry, getting rid of illegible notes, and preventing important medical details from being forgotten. See What Are the Advantages of Electronic Health Records?, supra note 18. But these values are only realized when EHRs are "interoperable"—when they permit data from one provider's system to integrate with another's. See Interoperability, Off. of the Nat'l Coordinator for Health Info. Tech., https://www.healthit.gov/topic/interoperability [https://perma.cc/W5CF-6FH3] (last visited Mar. 28, 2020). While the Privacy Regime has failed to achieve interoperability, it is almost guaranteed when hospitals share the same EHR vendor. See, e.g., EHR Interoperability from Anywhere, Epic, https://www.epic.com/interoperability/ehr-interoperability-from-anywhere [https://perma.cc/ZFS2-SNDU] (last visited Sept. 8, 2020) (stating the Epic EHR platforms are interoperable with other providers' Epic EHRs); EHR Interoperability: Why Is It So Difficult?, Carecloud, https://www.carecloud.com/continuum/why-is-ehr-interoperability-so-difficult [https://perma.cc/UYC9-EPYZ] (last visited Jan. 18, 2020) ("By helping the smaller players, interoperability is a direct threat to the larger companies' business models. There aren't many businesses that want to facilitate their customer's ability to look for services at another company.").

36. Jackie Drees, KLAS: Epic, Cerner Dominate EMR Market Share, Becker's Health IT (Apr. 30, 2019), https://www.beckershospitalreview.com/ehrs/klas-epic-cerner-dominate-emr-market-share.html [https://perma.cc/Z8E7-WC76].

37. Jackie Drees, Epic Unveils Patient Data Research Initiative, New Software, Becker's Health IT (Aug. 29, 2019), https://www.beckershospitalreview.com/ehrs/epic-unveils-patient-data-research-initiative-new-software.html [https://perma.cc/5532-THRH].

38. Tanner, supra note 27, at 143; Record Retrieval, Cerner, https://www.cerner.com/solutions/record-retrieval [https://perma.cc/YVX7-YEXH] (last visited Sept. 26, 2020).

Fusion served only five million patients a month via its EHR platform[39]—relatively few compared to EPIC and Cerner—but sold batches of de-identified EHR data for up to $2 million.[40] While it is unknown how much money the larger firms reap, it is certainly far more.

Every day the business grows. UnitedHealth recently announced that it would launch its own EHR system.[41] Amazon, Berkshire Hathaway, and JP Morgan recently announced a joint healthcare venture, which will likely include a play for medical data.[42] Even Google, which had once abandoned the health-data business,[43] made an inauspicious return when the *Wall Street Journal* revealed that 150 Google employees had access to tens of millions of patient records.[44] Google has assured the public that it was not in violation of HIPAA.[45]

B. *The Privacy Regime: HIPAA and HITECH*

Section I.B.1 describes HIPAA's core patient-privacy safeguard, the Privacy Rule.[46] Section I.B.2 discusses how HITECH successfully subsidized the adoption of EHR systems and improved the quality and quantity of health data collected by medical records. Part II then uses this background to explain how these laws created the EHR data market.

1. *HIPAA's Privacy Rule*

a. *Protecting Health Information.* — The Privacy Rule establishes standards for healthcare providers, insurers, other "covered entities," and their "business associate[s]" with respect to the disclosure and transfer of health information—defined as "information . . . that: (1) Is created or received

---

39. About Practice Fusion, Practice Fusion, https://www.practicefusion.com/about [https://perma.cc/JZL9-3ENL] (last visited Dec. 15, 2019).

40. Tanner, supra note 27, at 141.

41. Rebecca Pifer, UnitedHealth to Launch 'Fully Integrated' EHR Next Year, Healthcare Dive (Oct. 18, 2018), https://www.healthcaredive.com/news/unitedhealth-to-launch-fully-integrated-ehr-next-year/540023 [https://perma.cc/54WJ-FSJJ].

42. See Susan Morse, Amazon, Berkshire, JPMorgan Venture Hires BCBS IT Exec Dana Safran for Data-Driven Position, Healthcare IT News (Nov. 20, 2018), https://www.healthcareitnews.com/news/amazon-berkshire-jpmorgan-venture-hires-bcbs-it-exec-dana-safran-data-driven-position [https://perma.cc/5SL6-GWRP] (observing that data analysis will be important to the healthcare venture).

43. See Tanner, supra note 27, at 120–21 (describing Google Health, created in 2008, which was supposed to create a single central databank for all EHRs).

44. Rob Copeland, Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans, Wall St. J. (Nov. 11, 2019), https://www.wsj.com/articles/google-s-secret-project-nightingale-gathers-personal-health-data-on-millions-of-americans-11573496790 (on file with the *Columbia Law Review*).

45. See Christina Farr, Congressional Democrats Demand Details on Google's Use of Patient Data by Dec. 6, CNBC (Nov. 18, 2019), https://www.cnbc.com/2019/11/18/google-ascension-health-data-deal-under-scrutiny-by-congressional-dems.html [https://perma.cc/5C62-E4HE] (reporting that Google denied using the collected patient data for advertising purposes in violation of HIPAA).

46. See The HIPAA Privacy Rule, HHS, https://www.hhs.gov/hipaa/for-professionals/privacy/index.html [https://perma.cc/P29W-T676] (last visited May 24, 2020).

by a [covered entity]; and (2) Relates to the . . . physical or mental health or condition of an individual; the provision of health care to an individual; or the . . . payment for the provision of health care to an individual."[47] Hospitals, clinics, other healthcare providers, insurance providers, and EHR vendors are all either "covered entities" or their "business associates" and must thus comply with the Privacy Rule.[48] Protected health information (PHI) is "a subset of health information, including demographic information collected from an individual," which "identifies the individual[] or . . . can be used to identify the individual."[49] Both paper medical records and EHRs are governed by the Privacy Rule because they contain PHI.[50]

The Privacy Rule's baseline principle is that covered entities may not transmit PHI to anyone other than the patient without either the patient's "consent" or "authorization," depending on the situation.[51] Patient authorization is a rigidly defined concept and requires signed "authorization forms," documents that may not be combined with any other document and must clearly describe both the PHI being disclosed and to whom it will be disclosed.[52] Patient consent, on the other hand, is undefined by HIPAA but clearly requires that patients be informed of a requested PHI disclo-

47. 45 C.F.R. § 160.103 (2019). The term "covered entities" encompasses "health plan[s,] . . . health care clearinghouse[s,] [and] . . . health care provider[s] who [transmit] any health information in electronic form in connection with a transaction." Id. The term "business associate" includes any entity that "creates, receives, maintains, or transmits protected health information" on behalf of a covered entity. Id. This term encompasses EHR vendors such as Epic and Cerner and myriad other health information technology firms, dubbed by the HIPAA Omnibus Rule (issued in 2013) as "Health Information Organizations," though that term remains undefined. Id.; see also Marla Durben Hirsch, HIPAA Business Associate Compliance by EHR Vendors Not Optional, FierceHealthcare (Apr. 9, 2020), https://www.fiercehealthcare.com/ehr/hipaa-business-associate-compliance-by-ehr-vendors-not-optional [https://perma.cc/9VSX-6VZ8].

48. See Business Associates, HHS, https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/business-associates/index.html [https://perma.cc/68SD-BSG7] (last visited May 24, 2020); Covered Entities and Business Associates, HHS, https://www.hhs.gov/hipaa/for-professionals/covered-entities/index.html [https://perma.cc/7XEP-FAYY] (last visited May 24, 2020).

49. 45 C.F.R. § 160.103.

50. See Summary of the HIPAA Privacy Rule, HHS, https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html#:~:text=A%20covered%20entity%20is%20permitted,health%20or%20health%20care%20operations [https://perma.cc/Q8NV-39HJ] (last visited May 24, 2020).

51. See 45 C.F.R. § 164.502(a) ("A covered entity . . . may not use or disclose protected health information except as permitted or required by [the Rules governing PHI]."); see also id. § 164.508.

52. See id. § 164.508(b)–(c). Mere consent may not be substituted for an authorization. Id. § 164.506(b). Authorizations are specifically required when covered entities wish to disclose psychotherapy notes, to use PHI for marketing purposes, or to sell PHI. Id § 164.508(a).

sure and be given the opportunity to refuse sharing their PHI.[53] Because acquiring consent or authorization from large numbers of patients is often difficult, HIPAA permits the "waiver of authorization."[54] Authorization waivers can be used to facilitate PHI-based research, but only if the investigator receives permission from a "privacy board" or an Institutional Review Board (IRB)—an often-used but rigorous process.[55]

The Privacy Rule's twin goals of protecting patient privacy and not unduly burdening the routine provision of healthcare—which necessitates disclosing PHI to various providers, insurers, and other entities—are difficult to balance.[56] Exceptions to patient consent and authorization requirements help PHI flow through to the many entities that need it, but they may consume the Rule itself.[57] Most importantly, covered entities may disclose PHI for purposes of "treatment, payment, or health care operations" without authorization or consent.[58] These exceptions could conceivably

---

53. See What Is the Difference Between "Consent" and "Authorization" Under the HIPAA Privacy Rule?, HHS, https://www.hhs.gov/hipaa/for-professionals/faq/264/what-is-the-difference-between-consent-and-authorization/index.html [https://perma.cc/D3WH-B4A6] (last visited May 24, 2020).

54. See 45 C.F.R. § 164.512. HIPAA's varied and specific requirements make acquiring individual authorizations difficult, as does seeking out all of the patients whose information is being disclosed. See Sharona Hoffman & Andy Podgurski, Balancing Privacy, Autonomy, and Scientific Needs in Electronic Health Records Research, 65 SMU L. Rev. 85, 119–23 (2012) (detailing the difficulties of continually obtaining informed consent); The HIPAA Privacy Rule: How May Covered Entities Use and Disclose Health Information, Priv. Rts. Clearing House, https://privacyrights.org/consumer-guides/hipaa-privacy-rule-how-may-covered-entities-use-and-disclose-health-information [https://perma.cc/2EG6-QE8W] (last updated July 14, 2014) ("'Authorization' is much more formal than 'consent' and involves a patient granting signed permission.").

55. See 45 C.F.R. § 164.512. The IRB review process is no simple rubber stamp; some studies find that the costs of seeking IRB review are too high relative to the risks posed to individual privacy and describe them as a barrier to research. See, e.g., Miria Kano, Christina M. Getrich, Crystal Romney, Andrew L. Sussman & Robert L. Williams, Costs and Inconsistencies in US IRB Review of Low-Risk Medical Education Research, 49 Med. Educ. 634, 634–37 (2015) (describing ubiquitous procedural blocks to research created by the IRB-approval process); see also infra section II.B.1 (describing the Common Rule, IRB Review, and the disparity in ethical regulations governing research conducted by academic and clinical researchers versus research conducted on de-identified data by BigMedTech firms).

56. See Summary of the HIPAA Privacy Rule, supra note 50 (describing the Rule's principal goal).

57. See Uses and Disclosures for Treatment, Payment, and Health Care Operations, HHS (Dec. 3, 2002), https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/disclosures-treatment-payment-health-care-operations/index.html [https://perma.cc/U4CZ-XA4F] (last updated Apr. 3, 2003) ("[T]he Rule generally prohibits a covered entity from using or disclosing . . . [PHI] unless authorized by patients, except where this prohibition would result in unnecessary interference with access to quality health care or with certain other important public benefits or national priorities.").

58. 45 C.F.R. § 164.506.

cover nearly all of the PHI generated during the provision of healthcare.[59] The Rule attempts to cabin these exceptions by requiring covered entities to disclose only the "minimum [quantity of PHI] necessary" for the particular data transaction.[60] But it is unclear how much bite this standard has. For example, it is necessarily true that the minimum amount of PHI a provider must disclose to its EHR vendor is all of the PHI contained in the EHR.[61] Moreover, insurers themselves tout their access to rich pools of EHR data, suggesting that the minimum necessary standard is no obstacle to amassing vast quantities of EHR data.[62]

b. *The De-Identification Dichotomy.* — This Note uses the term "de-identification dichotomy" to describe the sharp contrast between the complex regulatory regime governing PHI—including the need for patient authorization and consent, IRB approval processes, the exceptions to the rule of patient authorization, and the minimum necessary standard—and the absence of de-identified EHR-data regulations. In order for PHI to become de-identified it must be stripped of names, addresses, birth dates, admission and discharge dates, telephone numbers, email addresses, social security numbers, plan provider numbers, biometric identifiers, and full face photographs.[63] With these fields removed, data can be freely used and transferred without patient or provider knowledge, IRB supervision, or any other limitation.[64]

---

59. See Uses and Disclosures for Treatment, Payment, and Health Care Operations, supra note 57 (listing extensive applications of these exceptions including "[b]illing and collection activities," "legal[] and auditing services," "[b]usiness management and general administrative activities," and much more).

60. 45 C.F.R. § 164.514(d).

61. Covered entities must limit the disclosure of PHI but are permitted to rely "on a requested disclosure as the minimum necessary . . . when . . . [t]he information is requested by another covered entity." Id. § 164.514(d)(3)(iii). Despite this standard, insurers sometimes demand authorization from patients for full PHI disclosure as a condition to coverage. See, e.g., Abigail English, Robin Summers, Julie Lewis & Clare Coleman, Nat'l Fam. Plan. & Reprod. Health Ass'n, Confidentiality, Third-Party Billing, & the Health Insurance Claims Process: Implications for Title X, at 7–8 (2015), https://www.nationalfamilyplanning.org/file/confidential–covered/ConfidentialandCovered_WhitePaper.pdf [https://perma.cc/H42E-GXSZ] (quoting Blue Cross & Blue Shield of Ala., Insert for 320 Plan, MKT-320 (7-2008), http://www.aldoi.gov/PDF/Consumers/320%20Plan%20%20policy.pdf [https://perma.cc/7Q6A-R98V] (last visited Sept. 4, 2020)).

62. For instance, United Health Group's Optum has "216 million lives of [EHR-derived] clinical and claims data." See Data: Establish the Foundation for Better Outcomes, supra note 29.

63. 45 C.F.R. § 164.514(b). To be considered de-identified, data cannot be readily re-identifiable, as determined by an expert. Id. Covered entities are entitled to transfer PHI to business associates for the purpose of creating de-identified data regardless of the data's ultimate end user. Id § 164.514(e)(3)(ii).

64. See id. § 164.502(d); see also Hoffman & Podgurski, supra note 54, at 139 ("But de-identified information is not covered by the privacy regulations, so the Privacy Rule does not entitle patients to any notice regarding research uses of data without identifiers." (footnotes omitted)).

The purpose of drawing attention to this stark contrast is explored further in section II.A, but chiefly it demonstrates the extent to which HIPAA balances privacy and the provision of healthcare, without directly addressing the flow of data for secondary uses. To be clear, the de-identification dichotomy is an important part of promoting valuable research without compromising patient privacy.[65] But as it is written, HIPAA does not account for its interaction with other key laws. Consequently, the de-identification dichotomy has failed to promote research while primarily benefiting BigMedTech.[66]

2. *Building a National Health Informatics Infrastructure*. — In 2009, Congress passed HITECH as part of the American Recovery and Reinvestment Act; subsidizing EHRs, it was thought, would both stimulate the economy and improve healthcare outcomes.[67] For the purposes of this Note, HITECH can be broken down into two broad components. First, it subsidized and strongly encouraged adoption of EHR technology, leading to a rapid rise in adoption and use rates. Second, it sought to make EHRs "interoperable," allowing "health data [to] flow freely, privately, and securely to the places where they are needed."[68] While HITECH's purpose was to establish "a nationwide health information technology infrastructure,"[69] this Note considers it a key part of the Privacy Regime because it balanced its principal goals against patient privacy (most notably via the Breach Notification Rule) but did not seek to directly shape the flow of EHR data for secondary uses.[70]

---

65. See Hoffman & Podgurski, supra note 54, at 104–05, 130–31 (explaining the value of de-identification for protecting patient privacy and the possibilities of using de-identified EHRs for research); Marc A. Rodwin, Patient Data: Property, Privacy & the Public Interest, 36 Am. J.L. & Med. 586, 589 (2010) (calling for the creation of a national database of de-identified EHRs in order to promote research while protecting patient privacy).

66. See infra note 189 and accompanying text (reviewing various works of scholarship proposing solutions to the problem of clinical researchers lacking access to large EHR databases).

67. See 42 U.S.C. § 300jj-11(b) (2018) (identifying HITECH's goals, which include: "(1) ensur[ing] . . . health information is secure . . . ; (2) improv[ing] health care quality . . . ; (3) reduc[ing] health care costs . . . ; ([4]) improv[ing] the coordination of care and information among [healthcare providers] . . . ; [and] ([5]) facilitat[ing] health and clinical research"); see also David Blumenthal, Launching HITECH, 362 New Eng. J. Med. 382, 382 (2010) ("The provisions of the HITECH Act are best understood not as investments in technology per se but as efforts to improve the health of Americans and the performance of their health care system."); Stephen P. Nash, Holme Roberts & Owen LLP, Stimulus Bill Contains the HITECH Act and Other Key HIT Provisions, Lexology 1–3 (2009), https://s3.amazonaws.com/documents.lexology.com/fc9d9b11-456e-424a-825e-0eeef830c194.pdf?AWSAccessKeyId=AKIAVYILUYJ754JTDY6T [https://perma.cc/KN4N-FRX3] (discussing HITECH's place in the broader plans to revive the American economy after the 2008 recession).

68. Blumenthal, supra note 67, at 382.

69. 42 U.S.C. § 300jj-11(b).

70. The Breach Notification Rule requires disclosure when a covered entity disseminates PHI outside of HIPAA's boundaries. See 45 U.S.C. § 17932 (2018). HITECH also

a. *Encouraging EHR Adoption.* — Pursuant to HITECH, the Centers for Medicare and Medicaid Services (CMS) (part of HHS) promulgated the three-stage Meaningful Use (MU) regulations subsidizing the adoption of EHRs.[71] In exchange for adopting and actively using a CMS-certified EHR platform, healthcare providers received increased Medicare and Medicaid reimbursements (that is, payment in excess of the usual reimbursement rate for a given medical procedure).[72] All told, MU payments to providers have exceeded $36 billion.[73] MU Stage One incentivized entry of PHI into EHR systems.[74] Stage Two set over twenty goals for participating providers, including (among others) use of EHRs to record medication, laboratory and radiology test results, demographic information, patient vital signs, smoking status, drug prescriptions, referrals to specialists, and summaries of care.[75] Stage Three expanded on the Stage Two goals, but its implementation coincided closely with the passing of the Medicare Access and CHIP

---

requires that the Privacy Rule and Security Rule's provisions (the latter covers the physical and digital safeguards intended to prevent unintended and unauthorized dissemination of PHI) and all compliance-related penalties apply equally to business associates and covered entities. See id. §§ 17931, 17934; see also Frank Pasquale, Grand Bargains for Big Data: The Emerging Law of Health Information, 72 Md. L. Rev. 682, 710 (2013) [hereinafter Pasquale, Grand Bargains] ("The law focuses on incentives to improve . . . population health, all while protecting privacy, confidentiality and security." (citing Nicholas P. Terry & Leslie P. Francis, Ensuring the Privacy and Confidentiality of Electronic Health Records, 2007 U. Ill. L. Rev. 681, 691–96)).

71. See Sharona Hoffman, Electronic Health Records and Medical Big Data: Law and Policy 38, 42 (2016). These subsidies varied according to whether the provider was an "eligible professional" or an "eligible hospital" and whether they participated in the Medicare or Medicaid program (which are different but function similarly). See id. at 39–41. EHR certification is overseen by HHS's Office of the National Coordinator for Health Information Technology (ONC), which tests EHR design, usability, task-completion speed, data quality, data collection type, and other metrics. Id. at 47–49.

72. See 42 C.F.R. § 495.2(a) (2019) (explaining the purpose of the MU "payment incentives under Medicare Part B [and Medicaid] for eligible professionals who adopt and meaningfully use certified [EHR] technology").

73. See Lucia Savage, Martin Gaynor & Julia Adler-Milstein, Digital Health Data and Information Sharing: A New Frontier for Health Care Competition?, 82 Antitrust L.J. 593, 599 (2019) (explaining that HITECH provided over $36 billion in incentive payments for physicians and hospitals to adopt and meaningfully use (as specified by CMS "Meaningful Use" criteria) software (with functions prescribed by ONC) to keep track of their patients' medical care through EHRs); see also Joseph D. Szerejko, Note, Reading Between the Lines of Electronic Health Records: The Health Information Technology for Economic and Clinical Health Act and Its Implications for Health Care Fraud and Information Security, 47 Conn. L. Rev. 1103, 1112 (2015) ("Incentive payments under the HITECH Act are only available to the providers that demonstrate meaningful use of EHRs.").

74. See 42 C.F.R. § 495.2(d). The types of PHI are further broken down to include entry of drug prescriptions and drug allergies, for example. Id.

75. See Hoffman, supra note 71, at 43. Stage Two also required submission of "clinical quality measures . . . designed to determine and track the quality of healthcare services" such as "the percentage of patients fifty to seventy-five years of age who" were screened for colorectal cancer. Id. at 44–45.

Reauthorization Act of 2015 (MACRA).[76] MACRA established a separate but overlapping incentive program, the Merit-Based Incentive Payment System (MIPS), charged with promoting interoperability, reducing the cost of care, and improving care quality.[77]

HITECH and MU were astoundingly successful in encouraging adoption and use of EHR systems. EHR penetration rates among office-based physicians increased from twenty percent in 2004 to eighty-six percent in 2017.[78] For hospitals, penetration increased from less than ten percent in 2008 to ninety-six percent in 2017.[79] Finally, there is strong empirical evidence that this rapid rate of adoption (of a complex and often maligned technology[80]) would not have occurred without HITECH's incentives.[81]

b. *Interoperability.* — A second major goal of HITECH, MACRA, MU, and MIPS (and the primary purpose of Congress's latest effort at improving the nation's health information technology infrastructure, the 21st Century Cures Act[82]) was improving the movement of data across providers, insurers, EHR platforms, and time—commonly referred to as EHR "interoperability."[83] Interoperability enables patients to freely travel between providers with all of their medical information (thus reducing costly duplicative labor and medical errors) and providers to switch between EHR vendors without losing patient records.[84] On a broader level, it is hoped that achieving interoperability will foster a national healthcare

---

76. Medicare Access and CHIP Reauthorization Act of 2015, Pub. Law No. 114–10, 129 Stat. 87 (codified in scattered sections of 42 U.S.C.).

77. See MIPS Overview, Ctrs. for Medicare & Medicaid Servs., https://qpp.cms.gov/mips/overview [https://perma.cc/66M2-T9E4] (last visited Mar. 28, 2020); Jennifer Morency, Meaningful Use 3 or MACRA? Same Same, but Different, Hello Health (July 1, 2018), https://hellohealth.com/blog/meaningful-use-3-or-macra-same-same-but-different [https://perma.cc/2NYX-ZMVG].

78. See Quick Stats, Off. of the Nat'l Coordinator for Health Info. Tech., https://dashboard.healthit.gov/quickstats/quickstats.php [https://perma.cc/7HG2-RN5K] (last updated June 17, 2019).

79. Non-Federal Acute Care Hospital Electronic Health Record Adoption, Off. of the Nat'l Coordinator for Health Info. Tech., https://dashboard.healthit.gov/quickstats/pages/FIG-Hospital-EHR-Adoption.php [https://perma.cc/3YCD-7LNW] (last visited Sept. 9, 2020).

80. See Atul Gawande, Why Doctors Hate Their Computers, New Yorker (Nov. 12, 2018), https://www.newyorker.com/magazine/2018/11/12/why-doctors-hate-their-computers (on file with the *Columbia Law Review*).

81. See, e.g., Julia Adler-Milstein & Ashish K. Jha, HITECH Act Drove Large Gains in Hospital Electronic Health Record Adoption, 36 Health Affs. 1416, 1421–22 (2017) ("We found that HITECH drove annual gains in EHR adoption of 8 percentage points in the five years after implementation . . . . [T]his result is dramatic and suggests that HITECH can serve as a model for other countries . . . .").

82. 21st Century Cures Act, Pub. L. No. 114–255, 130 Stat. 1033 (2016) (codified as amended in scattered sections of 42 U.S.C.).

83. See Interoperability, supra note 35 (discussing how these various stakeholders interact to encourage interoperability).

84. See Hoffman, supra note 71, at 54.

system that uses EHR data to continuously improve the provision and deployment of care, ultimately improving the health of all Americans.[85]

The government has been less successful on the interoperability front. HHS's Office of the National Coordinator for Health Information Technology (ONC) implements the various governmental efforts at achieving interoperability.[86] Despite ONC's best efforts, scholars agree that interoperability lags far behind EHR penetration: "By 2015, . . . only 6% of health care providers could share patient data with other clinicians who use an EHR system different from their own."[87] It remains to be seen whether the 21st Century Cures Act will be successful where previous efforts have failed.

II. PRIVATIZING PRIVACY: HOW HIPAA AND HITECH GAVE AWAY THE EHR

The question posed by the information presented in Part I is: Did the Privacy Regime contribute to the formation of the EHR data market, and if so, how? Without any clear statements from BigMedTech firms regarding how they collect and use PHI, it is difficult to answer with absolute certainty. Nonetheless, publicly available information permits drawing some important conclusions. Section II.A proposes that the Privacy Regime has played a pivotal role in creating the EHR data market and continues to shape it by concentrating EHR data in the hands of a few large firms. Accepting this argument, the next logical question is: Why should it matter if EHR vendors and insurance providers make money on the side as long as they continue providing healthcare services while protecting patient privacy? Section II.B argues that the Privacy Regime not only allows BigMedTech to conduct medical research without any of the ethical or scientific supervision applicable to academic research but also prevents academic researchers from accessing this potentially lifesaving resource. Because these defects can be remedied without compromising patient privacy or EHR adoption, their recognition demands congressional action, discussed in Part III.

---

85. See Off. of the 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 1–3, https://www.healthit.gov/sites/default/files/ONC10yearInteroperabilityConceptPaper.pdf [https://perma.cc/BX9T-KYA5] (last visited Mar. 28, 2020).

86. See id.

87. See Miriam Reisman, EHRs: The Challenge of Making Electronic Data Usable and Interoperable, 42 Pharmacy & Therapeutics 572, 572–74 (2017); cf. Hoffman, supra note 71, at 55 ("Interoperability will not happen without regulatory intervention. Vendors are not motivated to promote interoperability because it will increase the likelihood that dissatisfied customers will jettison their existing EHR systems and switch to new ones made by different manufacturers.").

A. *The Privacy Regime Created the EHR Data Market and Funnels Data to BigMedTech*

Analyzing the EHR data market in light of the Privacy Regime yields three key findings providing compelling evidence that the existence of the EHR data market and the dominance of BigMedTech within that market are the direct result of the Privacy Regime.[88] First, section II.A.1 demonstrates that the EHR data market would not exist without both HITECH's subsidies and HIPAA's de-identification dichotomy. Next, section II.A.2 shows that HIPAA funnels data "upward" toward BigMedTech. Finally, section II.A.3 argues that HIPAA's de-identification dichotomy, the high cost of de-identification, and HITECH's failure to achieve interoperability ensure that EHR data cannot escape the HIPAA funnel.

1. *HIPAA and HITECH Created the EHR Data Market.* — As section I.A.1 demonstrates, HITECH increased both the scope and scale of data available for medical research. In the eight years after its passage, HITECH increased EHR penetration rates from approximately ten percent to ninety percent—thereby increasing the sheer number of analyzable medical records.[89] EHRs also increase the scope of available data by capturing more data points covering more measurements in each patient–doctor interaction—"cross-sectional" data capture—and by tracking patients across years or decades—"longitudinal" data capture.[90] It would be logistically and technically impossible to create comparably comprehensive medical record databases using paper records.[91]

The increased scope and scale of data capture were crucial to the formation of a robust data market. Clinical investigators have studied medical records for hundreds of years.[92] But the promise of big data for EHR-based

---

88. It is important to note at the outset of this analysis that the EHR data market, and the concentration within that market, owes its existence to several nonlegal economic factors, most prominently network effects. See supra notes 33–36 and accompanying text; see also infra notes 124–126 and accompanying text (discussing other such factors). Section II.A, however, shows that without the Privacy Regime, the EHR data market could not exist as it does today.

89. See supra notes 78–81 and accompanying text; see also Pascal Coorevits, Mats Sundgren, Gunnar O. Klein, Anne Bahr, Brecht Claerhout, Christel Daniel, Martin Dugas, Danielle Dupont, Andreas Schmidt, Peter Singleton, Georges De Moor & Dipak Kalra, Electronic Health Records: New Opportunities for Clinical Research, 274 J. Internal Med. 547, 549–50 (2013) ("Indeed, a very dramatic recent increase [in EHR adoption] in the USA has been largely due to government financial incentives for EHRs with 'meaningful use' criteria.").

90. See supra notes 16–21 and accompanying text.

91. Again, several BigMedTech firms exploit hundreds of millions of patient medical records. See supra section I.A.2. Without EHRs, data would have to be manually transferred from paper records to electronic databases before any analysis could take place.

92. See Richard F. Gillum, From Papyrus to the Electronic Tablet: A Brief History of the Clinical Medical Record with Lessons for the Digital Age, 126 Am. J. Med. 853, 854 (2013) ("New research methods such as the numerical method of Pierre Louis (1787–1872)

research lies in leveraging millions, not hundreds or thousands, of records.[93] Scholar Shoshana Zuboff describes big data as "the ultimate tapeworm" because its analytical power depends entirely on the amount of data available for processing.[94] Data markets, and the firms that dominate them, rely on massive volumes of information to feed this tapeworm and must maximize data capture across every possible dimension.[95] Thus, HITECH created the raw materials necessary for the medical-record data market's existence.

HITECH gave BigMedTech firms a valuable resource to exploit, but without HIPAA there would not have been a legal framework within which to do so. Before HIPAA, a "patchwork" of potentially incompatible state laws governed the exchange of patient health information.[96] Generally, a clear and uniform regulatory regime is preferable for market participants.[97] But one could easily envision a world in which HIPAA simply pro-

---

could be applied to series of case histories to test hypotheses about disease causation or therapeutic efficacy, increasing the value of archives of such histories in medical centers."); Dan R. Schlegel & Gregorie Ficheur, Secondary Use of Patient Data: Review of the Literature Published in 2016, 26 Y.B. Med. Informatics 68, 68 (2017) (finding that medical record data are often reused for research by the institution that first collects it).

93. See supra notes 20–23 and accompanying text; see also Barbara J. Evans, Barbarians at the Gate: Consumer-Driven Health Data Commons and the Transformation of Citizen Science, 42 Am. J.L. & Med. 651, 655 (2016) [hereinafter Evans, Barbarians at the Gate] (noting that today's medical challenges "require access to very large-scaled data resources—sometimes, data for . . . millions of individuals . . . . The most valuable data resources are deeply descriptive[,] . . . [and contain] a rich array of genomic and other diagnostic test results, clinical data" and more); Pasquale, Grand Bargains, supra note 70, at 683 ("Quantitative analysis of large information sets ('big data') has spurred scientific and business breakthroughs. Better collection and analysis of health data may save lives, cut costs, and expand access to care."); cf. Rodwin, supra note 65, at 586 ("Tapping data from patient records would make possible similar evaluations at much lower costs, yield continually updated information, and facilitate rapid learning. It would provide information on populations and variables not included in clinical trials.").

94. Shoshana Zuboff, The Age of Surveillance Capitalism 95 (2019); see also Katherine S. Button, John P.A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S.J. Robinson & Marcus R. Munafò, Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience, 14 Nature Revs. Neuroscience 365, 365–76 (2013) (showing that increased statistical power improves both the accuracy and significance of statistical predictions).

95. See Zuboff, supra note 94, at 201 (describing dominant firms' leveraging of "economies of scope" as requiring not only a vast number of individual data subjects or records, but variation and depth in the characteristics measured).

96. See Why Is the HIPAA Privacy Rule Needed?, HHS, https://www.hhs.gov/hipaa/for-professionals/faq/188/why-is-the-privacy-rule-needed/index.html [https://perma.cc/SD6S-QKVJ] (last visited Sept. 4, 2020).

97. See Alan McQuinn & Daniel Castro, The Case for a U.S. Digital Single Market and Why Federal Preemption Is Key, Info. Tech. & Innovation Found. (Oct. 7, 2019), https://itif.org/publications/2019/10/07/case-us-digital-single-market-and-why-federal-preemption-key [https://perma.cc/JFD2-5VVY] ("States and localities have created barriers to digital commerce through overlapping and conflicting rules, including in areas of data privacy and

hibited *any* sale of medical data or *any* research without patient consent. Instead, HIPAA established a sharp line: On one hand, identifiable data cannot be sold, and their movement and use are governed by a complex regulatory regime—while on the other hand, de-identified data may be analyzed and sold with almost no regulation.[98] By clearly limiting itself to regulating identifiable medical data, HIPAA opened the door to a robust market for de-identified EHR data.[99]

2. *HIPAA Gives BigMedTech an EHR Data Pipeline.* — The Privacy Regime also provides BigMedTech with the mechanism for acquiring EHR datathrough the Privacy Rule's treatment, payment, and operations exceptions to authorization and consent requirements. To be clear, these exceptions are broad: They all "require use and disclosure of protected health information, [and] are [all] essential to the effective operation of the health care system."[100] By definition, all PHI collection is related to patient treatment.[101] An insurer has several legitimate avenues for aggregating PHI, such as eligibility and coverage determinations, billing and collection activities, medical necessity reviews, credit bureau disclosures, and "certain administrative, financial, legal, and quality improvement activities . . . that are necessary to run its business."[102] And while EHR vendors are not clearly referenced in the Rule or in any guidance, their centrality to gathering, storing, and moving PHI requires that they have access to the entire EHR.[103]

The Privacy Rule's exceptions are laid atop economic dynamics in the healthcare industry, and together they form a large-capacity data pipeline

---

net neutrality, and Congress has failed to stop states from erecting these policies."); cf. Alan Schwartz, Statutory Interpretation, Capture, and Tort Law: The Regulatory Compliance Defense, 2 Am. L. & Econ. Rev. 1, 20–22 (2000) (arguing that preemption of state product safety laws is preferable to a market governed by differing state laws).

98. See supra section I.B.1. In particular, PHI cannot be sold without a patient-signed document detailing what PHI is being sold and to whom. Each further sale of that same PHI would require another signed authorization form. 45 C.F.R. § 164.502(a)(5)(ii) (2019).

99. This is especially true given that medical data need not be identifiable to be valuable. See Rodwin, supra note 65, at 609 ("Aggregate patient data is valuable for purposes that do not require identifying individuals[] . . . .").

100. See 45 C.F.R. § 164.506; Uses and Disclosures for Treatment, Payment, and Health Care Operations, supra note 57.

101. See 45 C.F.R. § 164.501 (defining "treatment" as including "consultation between health care providers relating to a patient" during which PHI is gathered in the EHR); see also How May the HIPAA Privacy Rule's Minimum Necessary Standard Apply to Electronic Health Information Exchange Through a Networked Environment, HHS (Dec. 15, 2008), https://www.hhs.gov/hipaa/for-professionals/faq/545/how-may-hipaas-minimum-necessary-standard-apply-to-electronic-information/index.html [https://perma.cc/JL5T-U4ZS] (last updated July 26, 2013) (stating that PHI transfers between providers for treatment purposes are not subject to the minimum necessary standard).

102. Uses and Disclosures for Treatment, Payment, and Health Care Operations, supra note 57.

103. Cf. supra notes 59–62 and accompanying text (discussing the vagueness of the "minimum necessary" standard and the breadth of the Privacy Rule's authorization exceptions).

for BigMedTech firms. There were thirty-six million admissions in 2018 to the country's six thousand hospitals, and ninety million patients visited its eight thousand urgent care centers.[104] In turn, thanks to network effects and economies of scale,[105] well over half of the EHR market is captured by Epic and Cerner and four insurance companies dominate healthcare payment.[106] Because big data analysis requires massive quantities of data, these firms are incentivized to gather as much of the EHR data flowing through their networks as possible.[107] To the extent that providers could push back on BigMedTech's use of EHR data, it is not clear why they would, given the comparably meager volume and value of their data.[108] Perhaps more importantly, while almost all hospitals must use some data to report quality measures to the federal government,[109] clinical and academic research simply are not critical to the mission of providing acute patient care.[110]

There is little official information on BigMedTech's data-aggregation practices. But reporting by Tanner and other journalists—corroborated by

104. Cheryl Alkon, What's Behind the Growth of Urgent Care Clinics?, Med. Econ. (Aug. 29, 2018), https://www.medicaleconomics.com/view/whats-behind-growth-urgent-care-clinics [https://perma.cc/PX8N-KVYM]; Fast Facts on U.S. Hospitals, 2020, Am. Hosp. Ass'n, https://www.aha.org/statistics/fast-facts-us-hospitals [https://perma.cc/XF9X-PLWG] (last updated Mar. 2020).

105. See supra notes 33–36 and accompanying text.

106. See supra note 36 and accompanying text. United Health Group has fifty million patients in its network, and Anthem has another forty million. See Shelby Livingston, UnitedHealth Revenue Grows with Medicare Members, OptumHealth Patients, Mod. Healthcare (Apr. 16, 2019), https://www.modernhealthcare.com/insurance/unitedhealth-revenue-grows-medicare-members-optumhealth-patients [https://perma.cc/7BBC-H9AG]; Stats and Facts, Anthem, https://www.antheminc.com/NewsMedia/FrequentlyRequestedMaterials/StatsFacts/index.htm [https://perma.cc/H5D5-S39L] (last visited Oct. 25, 2019).

107. See supra notes 16–23, 92–95 and accompanying text.

108. Given thirty-six million hospital admissions for six thousand hospitals, the average hospital sees approximately six thousand admissions per year, far fewer data points than those held by BigMedTech firms sitting at the top of the data collection pyramid. See Fast Facts on U.S. Hospitals, 2020, supra note 104.

109. See Mike Miliard, CMS Offers Quality Reporting Relief as Providers Battle COVID-19, Healthcare IT News (Mar. 23, 2020), https://www.healthcareitnews.com/news/cms-offers-quality-reporting-relief-providers-battle-covid-19 [https://perma.cc/ZN6Z-4ZE4] (stating that 1.2 million clinicians are required to report quality measures to CMS for reimbursement purposes).

110. About five percent of U.S. hospitals are considered to be academic medical centers (AMCs)—as opposed to community medical centers—but the research conducted at AMCs is credited with many of the most important medical breakthroughs in history. The Differences Between Academic and Community Medical Centers, Geo. Wash. Univ. Sch. of Bus., https://healthcaremba.gwu.edu/blog/the-differences-between-community-and-academic-medical-centers [https://perma.cc/3ZDD-JQD3] (last visited Apr. 12, 2020). Community medical centers have little secondary use for their EHR data beyond collecting the quality metrics they report to the government. See id.

the firms' websites—speaks to the large volume of data moving through the HIPAA-exceptions pipeline.[111]

3. *The Only Game in Town: How the Privacy Regime Prevents EHR Data from Escaping the Funnel.* — The corollary to the "funnel" dynamics described above is that the Privacy Regime also prevents EHR data from flowing outside of the funnel. This section first argues that the rules governing EHR-based clinical studies inhibit the large-scale aggregation of identifiable PHI. Second, because providers cannot sell PHI, firms and research institutions not in the business of treatment, payment, or healthcare operations cannot aggregate sufficient volumes of data. Third, the cost of de-identification and data management discourages most providers from withholding data from BigMedTech and from leveraging their EHR data themselves. Finally, the Privacy Regime's ongoing failure to achieve interoperability inhibits "horizontal" data flow—that is, data can easily flow "vertically" upward from providers to insurers, EHR vendors, and other firms, but not across providers or EHR systems—thereby increasing the value of HIPAA's exceptions for data aggregation.

BigMedTech's stores of de-identified EHR data are valuable in large part because the Privacy Regime prevents healthcare providers from assembling PHI databases on the scale required for big data analysis.[112] This makes sense from a privacy perspective: Allowing dozens or even hundreds of organizations to pool identifiable data for research is antithetical to the conceptions of individual privacy that underpin the Privacy

111. See Better Results for Your Employees When Caregivers Are Connected, Kaiser Permanente Bus., https://business.kaiserpermanente.org/insights/a-look-at-true-connectivity-in-health-care [https://perma.cc/X6XW-G8ZC] (last visited Sept. 5, 2020); In a Nutshell, Epic, https://www.epic.com/software#Cosmos [https://perma.cc/HNX8-U9FQ] (last visited Sept. 5, 2020); Record Retrieval, supra note 38; supra section I.A.2.

112. See supra section I.B.1.i (discussing HIPAA's complex regime covering the movement of PHI among covered entities); see also supra notes 16–23, 94–95 and accompanying text (discussing the promise of big data EHR analysis, which requires hundreds of thousands or millions of medical records). HIPAA does allow researchers to compile "limited data set[s]." See 45 C.F.R. § 164.514(e) (2019). Sixteen identifiers must be removed for data sets to be considered "limited" and must be accompanied by "data use agreements." Id. These agreements establish "the ways in which the information in the limited data set may be used and how it will be protected." How Can Covered Entities Use and Disclose Protected Health Information for Research and Comply with the Privacy Rule?, NIH, https://privacyruleandresearch.nih.gov/pr_08.asp [https://perma.cc/2U6X-NEQY] (last visited Apr. 11, 2020). Some limited data set requirements are relatively vague, for example: "[u]se [of] appropriate safeguards to prevent use or disclosure of the information other than as provided for by the data use agreement." 45 C.F.R. § 164.514(e)(4)(ii)(C)(*2*). Others are more demanding, requiring covered entities to "[n]ot use or further disclose the information other than as permitted by the data use agreement." Id. § 164.514(e)(4)(ii)(C)(*1*). Such limits inevitably render limited data sets less attractive for research, and more difficult to aggregate, than unregulated de-identified data.

Regime.[113] Clinical research (including PHI-based research[114]) is thus highly regulated under both HIPAA and the Common Rule, which applies to all human-subject research funded at least in part by any of twenty federal agencies (hence, a "common" rule).[115] The Common Rule permits HIPAA-covered entities to use EHR data for secondary research without the full IRB review applicable to studies that actively use human participants.[116] Nonetheless, most academic medical centers (AMCs)[117] require investigators to approach an IRB to confirm exemption before research can begin.[118] Thus, conducting population-wide research on millions of records would first require collaboration among dozens of IRBs, which commentators agree is logistically quite difficult.[119] It is far more efficient to first aggregate PHI through approved HIPAA channels before de-identifying the data and then commencing research. Gathering vast quantities of PHI before de-identification for a clinical study is not practically different from what BigMedTech firms do; the only difference is in the stated purpose.[120] Gathering it for research purposes entails the logistical obstacles described above, but none of those obstacles apply to gathering PHI for treatment, payment, or healthcare operations.[121]

---

113. See Summary of the HIPAA Privacy Rule, supra note 50 ("A major goal of the Privacy Rule is to assure that individuals' health information is properly protected . . . [and] to define and limit the circumstances in which an individual's protected heath [sic] information may be used or disclosed by covered entities."). Moreover, de-identifying and then pooling EHR data presents a significant technical challenge, which most organizations are simply not equipped to solve. See infra notes 126–131 and accompanying text (discussing the technical and financial challenges of EHR database management and de-identification).

114. For the purposes of research, the Privacy Rule defines a human subject as "a living individual about whom an investigator . . . [o]btains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens" where "[i]dentifiable private information is private information for which the identity of the subject is or may readily be ascertained by the investigator or associated with the information." 45 C.F.R. § 46.102(e)(1), (5).

115. Alda Yuan, Blurred Lines: The Collapse of the Research/Clinical Care Divide and the Need for Context-Based Research Categories in the Revised Common Rule, 74 Food & Drug L.J. 46, 48 (2019).

116. 45 C.F.R. § 46.104(d)(4); see also infra text accompanying note 158. More to the point, the government does not fund BigMedTech research and thus those firms need not comply with any of the Common Rule's requirements. See 45 C.F.R. § 46.101.

117. See supra note 110 (defining AMCs).

118. See infra section II.B (discussing in greater detail the legal and extralegal constraints imposed on academic researchers that are absent in for-profit research).

119. See, e.g., E. Andrew Balas, Marlo Vernon, Farah Magrabi, Lynne Thomas Gordon & Joanne Sexton, Big Data Clinical Research: Validity, Ethics, and Regulation, 216 Stud. Health Tech. & Informatics 448, 450 (2015) ("Unfortunately, variable interpretations and a lack of coordination among multi-site IRBs creates a challenging health research environment.").

120. See supra sections I.A.2–.B.1, II.A.2 (discussing the state of the EHR data market and the treatment, operations, and payment mechanisms used by BigMedTech to gather this data).

121. See supra notes 100–111 and accompanying text.

The Privacy Regime also makes any kind of data aggregation nearly impossible for nonhealthcare firms. If providers were permitted to sell PHI, other firms could potentially compile large databases and compete with BigMedTech.[122] Though even then, they would face the disadvantage of paying for data that BigMedTech firms gather freely during other profit-generating activities.[123] Thus, regardless of the prohibition on PHI sales, the PHI funnel grants BigMedTech firms a nearly insurmountable advantage for gathering PHI relative to outside firms or clinical researchers.

As discussed above, most U.S. hospitals do not conduct clinical research, and therefore may not be attuned to the possibility of using their data for such purposes[124]—but even if they were, de-identification and EHR data management makes contributing data prohibitively costly. Manual HIPAA-compliant de-identification is expensive—though necessary because physician's notes often contain patient-identifying details.[125] But automated processes under development are likely to be expensive as well. They require the application of algorithms trained on EHRs where identifying information has been manually flagged by researchers.[126] And even after de-identification, converting EHR data to an analyzable format requires processing the structured (temperatures, weights, dosages, mortality, etc.) and unstructured (clinical notes, radiological images, etc.) data into a standardized form.[127] Both steps are complex, but the latter is especially so, requiring the application of advanced (and constantly evolving) machine-learning techniques known as "natural language processing."[128] Providers would then need to coordinate with their peers to connect the

---

122. Cf. 45 C.F.R. § 164.502(a)(5)(ii) (2019) (forbidding the sale of PHI outside of very limited circumstances). Moreover, this Note does not dispute that permitting the sale of PHI would unacceptably compromise personal privacy and autonomy.

123. See supra sections I.B.1, II.A.2.

124. See supra note 110 and accompanying text.

125. See Stephane Meystre, F. Jeffrey Friedlin, Brett R. South, Shuying Shen & Matthew H. Samore, Automatic De-Identification of Textual Documents in the Electronic Health Record: A Review of Recent Research, BMC Med. Rsch. Methodology, Aug. 2010, at 1, 2 ("Dorr et al. have evaluated the time cost to manually de-identify narrative text notes . . . and concluded that it was time-consuming and difficult to exclude all PHI required by HIPAA."); Paul Govern, Report Seeks to Streamline EHR De-Identification, Vand. Univ. Med. Ctr.: VUMC Rep. (Apr. 4, 2019), https://news.vumc.org/2019/04/04/report-seeks-to-streamline-ehr%E2%80%88de-identification [https://perma.cc/7E5L-NH8T] (discussing the currently expensive process of applying machine learning to medical records to de-identify physician notes). But see Stephane M. Meystre, De-Identification of Unstructured Clinical Data for Patient Privacy Protection, *in* Medical Data Privacy Handbook 697, 698 (Aris Gkoulalas-Divanis, Grigorios Loukides eds., 2015) ("Automated approaches based on Natural Language Processing (NLP) have been implemented and evaluated, allowing for much faster de-identification than manual approaches.").

126. Govern, supra note 125.

127. See Elizabeth S. Chen & Indra Neil Sarkar, Mining the Electronic Health Record for Disease Knowledge, *in* Biomedical Literature Mining 269, 269–71 (Vinod D. Kumar & Hannah Jane Tipney eds., 2014) (describing the text-mining process, which includes "data selection, preprocessing, transformation, data mining, and interpretation/validation").

128. See id.

de-identified data sets, an added expense of time and money.[129] Finally, data scientists must apply advanced algorithms to the data to detect patterns and test hypotheses.[130] These expenses, in addition to costly EHR vendor fees,[131] further disincentivize providers from withholding EHR data from BigMedTech as they negotiate for EHR services or insurance reimbursements.

Finally, the Privacy Regime's failure to achieve true interoperability further inhibits data aggregation by organizations outside of BigMedTech. HITECH and similar laws were supposed to enable the horizontal movement of data—again, data flowing with patients to multiple providers, insurers, and EHR vendors. But the ongoing inability to achieve this goal means data primarily moves vertically—upward from patients to providers to BigMedTech firms—via the HIPAA funnel.[132] Here, the law cannot bear all of the blame. Interoperability is an extraordinarily challenging technical problem facing many headwinds including: skepticism from providers, the high cost of upgrading EHR systems, incompatibly customized EHR systems (to meet the needs of differing institutional workflows), shifting regulatory requirements, and more.[133] Even if interoperability were achieved, it would not instantaneously create large shared databases; researchers have many nontechnical disincentives to sharing data.[134] Nonetheless, many believe that true interoperability could empower researchers to share, aggregate, and study EHR data in collaboration.[135]

---

129. In contrast, BigMedTech firms continuously acquire massive volumes of data and do not need to coordinate with their peer firms. See supra section I.A.2 (discussing how much data various BigMedTech firms hold); see also supra notes 116–121 and accompanying text (discussing the challenge of coordinating various IRBs for the purpose of building shared research resources).

130. See, e.g., Joan A. Casey, Brian S. Schwartz, Walter F. Stewart & Nancy E. Adler, Using Electronic Health Records for Population Health Research: A Review of Methods and Applications, 37 Ann. Rev. Pub. Health 61, 65 (2016) (describing the challenges of preparing EHR data for research).

131. See Unpacking Hospitals' EHR Implementation Costs: What's Behind the Million-Dollar Price Tags?, Becker's Health IT, https://www.beckershospitalreview.com/healthcare-information-technology/unpacking-hospitals-ehr-implementation-costs-what-s-behind-the-million-dollar-price-tags.html [https://perma.cc/U3FE-RZXB] (last visited May 11, 2020).

132. See supra note 87 and accompanying text (finding that only six percent of EHR systems are thought to be interoperable in any meaningful way).

133. See Am. Hospital Ass'n Interoperability Advisory Grp., Achieving Interoperability that Supports Care Transformation 9 (2015), https://www.aha.org/system/files/2018-10/1507-iagreport.pdf [https://perma.cc/B8KY-WLAT]; Chris Yager, Achieving Interoperability in Healthcare, Becker's Health IT (May 31, 2018), https://www.beckershospitalreview.com/healthcare-information-technology/achieving-interoperability-in-healthcare.html [https://perma.cc/X6PZ-J5TZ]; Reisman, supra note 87, at 575.

134. See Willem G. van Panhuis, Proma Paul, Claudia Emerson, John Grefenstette, Richard Wilder, Abraham J. Herbst, David Heymann & Donald S. Burke, A Systematic Review of Barriers to Data Sharing in Public Health, 14 BMC Pub. Health 1144, 1146–49 (2014) (reviewing the literature to find twenty distinct barriers only seven of which are technical challenges).

135. See Coorevits et al., supra note 89, at 547–48.

Meanwhile, its absence has no doubt helped clear the field for BigMedTech firms, which sit atop the Privacy Regime's powerful data-aggregation apparatus.

All this said, other technology firms have attempted to penetrate the EHR data market. Google is the most obvious—its partnerships with Ascension, Mayo Clinic, and other large hospital systems provide access to valuable stores of high-quality data.[136] But while studies on thousands of patient records are quite valuable, the promise of EHR-based clinical research lies in achieving massive scale.[137] On this front Google has been frustrated, having been rebuffed by Cerner despite offering up to $250 million in discounts for its data-storage services.[138] More recently, Google and its peer firms may have scored a victory in a new interoperability rule promulgated by CMS in May 2020.[139] The rule allows patients to integrate their EHR data with third-party health and fitness applications on individual users' various devices.[140] This could give firms like Google, which recently acquired Fitbit with its twenty-eight million users,[141] the opportunity to amass vast quantities of health data. Epic vehemently opposes the rule; it sent letters to many prominent healthcare providers urging opposition to the rule and promises to sue to block its implementation.[142] It

136. See Rob Copeland, Dana Mattioli & Melanie Evans, Inside Google's Quest for Millions of Medical Records, Wall St. J. (Jan. 11, 2020), https://www.wsj.com/articles/paging-dr-google-how-the-tech-giant-is-laying-claim-to-health-data-11578719700 (on file with the *Columbia Law Review*); Mike Miliard, Mayo Clinic, Google Launch Major New 10-Year Partnership, Healthcare IT News (Sept. 11, 2019), https://www.healthcareitnews.com/news/mayo-clinic-google-launch-major-new-10-year-partnership [https://perma.cc/QGX2-3HQU] ("Mayo Clinic and Google are embarking on a decade-long strategic partnership with advanced cloud computing and AI-powered analytics at its core. Together, the two giants seek to spur huge new innovations for care delivery at the health system and beyond.").

137. See supra notes 20–26, 93–95 and accompanying text.

138. See Copeland et al., supra note 136 ("Google had a bigger goal in pushing for the [Cerner] deal than dollars and cents: a way to expand its effort to collect, analyze and aggregate health data on millions of Americans."). Cerner ultimately signed a similar deal with Amazon. Id.

139. CMS Interoperability and Patient Access Final Rule, 45 C.F.R. § 170.215 (2019).

140. See Jennifer Bresnick, CMS Sparks Mixed Reactions with Interoperability, Data Blocking Rules, Health IT Analytics (Feb. 15, 2019), https://healthitanalytics.com/features/cms-sparks-mixed-reactions-with-interoperability-data-blocking-rules [https://perma.cc/N3CP-LRM7]; see also Christopher Jason, Epic Leads Almost 60 Health Systems Against Interoperability Rule, EHR Intel. (Feb. 6, 2020), https://ehrintelligence.com/news/epic-leads-almost-60-health-systems-against-interoperability-rule [https://perma.cc/N74H-85G5] (noting that the rule "supports patients accessing and sharing their own electronic health information via mobile apps").

141. Brian Heater, Google Is Acquiring Fitbit for $2.1 Billion, TechCrunch (Nov. 1, 2019), https://techcrunch.com/2019/11/01/google-is-acquiring-fitbit [https://perma.cc/AG9N-VBPY].

142. See Dane Finley, Epic May File a Lawsuit Against the US Department of Health and Human Services Over Its Data-Sharing Rules, Bus. Insider (Jan. 28, 2020), https://www.businessinsider.com/epic-may-sue-department-of-health-human-services-2020-1 [https://perma.cc/F79M-GF5N]; Jason, supra note 140.

remains to be seen whether this rule will result in interoperability that benefits either patients or medical research. Epic maintains that it is fighting the rule out of concern for patient privacy.[143]

B. *The Price of Privacy: Unsupervised Research and Unanswered Questions*

This section shows that by funneling EHR data to BigMedTech, the Privacy Regime creates two distinct and incompatible sets of rules for medical research. One set applies to the academic world: It requires IRB approval, peer review, and publication. The other set applies to BigMedTech firms and permits research that is completely unregulated, vulnerable to bias, and unpublished. It is important to remember from the outset that BigMedTech owes its ability to sell EHR-based research to a legal regime—for that reason, lawmakers must be attuned to the Privacy Regime's potentially harmful consequences.[144] First, section II.B.1 argues that this regime fails to place BigMedTech's research under the ethical scrutiny to which the exact same research would be submitted if it were being conducted by academic researchers. Then, section II.B.2 argues that this research may be methodologically flawed in the absence of academic participation because it is conducted by non-clinicians and without peer review or publication.

The problem is not that each and every study conducted by BigMedTech firms harms patients—indeed, many may be beneficial. The problem is that the Privacy Regime fails to systematically maximize research benefits while minimizing possible harms, which is a fundamental principle of medical research ethics.

1. *The Absence of Ethical Supervision Applicable to Clinical Research Is Anathema to Medical Ethics.* — The medical establishment has long recognized the need to "foster . . . innovative research [and] . . . expand the knowledge base in medical and associated sciences" while mitigating possible harm to research subjects.[145] In response to Nazi human experimentation, the Nuremberg Code of 1947 announced the world's first set of

---

143. Jason, supra note 140 ("Although Faulkner says the company supports patient access to their data, she states the rule will result in app makers gaining access to patient data without consent.").

144. This Note does not review the potential privacy-related harms (such as re-identification or the risk of a security breach) posed by such research. See Hoffman & Podgurski, supra note 54, at 102–07. Instead, it attempts to demonstrate that even a hypothetically perfect privacy regime will systematically fail to account for other harms without imposing ethical review.

145. See Mission and Goals, NIH, https://www.nih.gov/about-nih/what-we-do/mission-goals [https://perma.cc/XZY9-KUP8] (last visited Jan. 15, 2020). Clinical research is the class of activities designed to test hypotheses, "permit conclusions to be drawn, and . . . contribute to generalizable knowledge." Dep't of Health, Educ. & Welfare, Off. of the Sec'y, The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, HHS (Apr. 18, 1979), https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf [https://perma.cc/P5N2-YFFD] [hereinafter the Belmont Report].

ethical principles for medical research.[146] In the mid-1970s widespread outrage followed the revelation that American researchers had inoculated hundreds of Black men with syphilis without their consent and without treating them for forty years in the Tuskegee study.[147] In response, the United States declared its commitment to the Nuremberg Code's standards in the famed Belmont Report, which announced the three foundational principles of ethical medical research: "respect [for] persons, beneficence, and justice."[148] The principle of "respect for persons" requires treating individuals as autonomous agents; it is embodied in the familiar medical and legal doctrine of informed consent.[149] The principle of "beneficence" incorporates two separate duties—first, "do not harm," and second, "maximize possible benefits and minimize possible harms"—and in practice requires that a study's benefits outweigh its risks.[150] Finally, "justice" requires fairness in the distribution of benefits and burdens in medical research.[151]

Generally, "if there is any element of research in an activity, that activity should undergo review for the protection of human subjects."[152] All investigators must safeguard the Belmont Principles in the course of any study, but IRBs are the Principles' formal safeguard.[153] They review any proposed research project's purpose and methodology for beneficence, informed consent or an appropriate informed consent waiver, and equitable subject selection.[154] They may also continuously monitor subject-safety where the study's design requires it.[155] Critically, even the least risky aca-

---

146. Joseph L. Breault, Protecting Human Research Subjects: The Past Defines the Future, 6 Ochsner J. 15, 15 (2006).

147. Id. at 15–16.

148. The Belmont Report, supra note 145; see also Jennifer Sims, A Brief Review of the Belmont Report, 29 Dimensions Critical Care Nursing 173, 173 (2010).

149. See the Belmont Report, supra note 145. "Informed consent" is the patient's right to "be given the opportunity to choose what shall or shall not happen to them" and mandates disclosure of sufficient information regarding the research procedure, purposes, risks, and benefits before the individual consents to the proposed intervention. Id.

150. See id. "Do not harm" is a separate duty from "minimize possible harms" in that the latter requires balancing possible (as opposed to known) harms against benefits before undertaking a particular procedure. Id. Harm can be "psychological," "physical," "legal," "social," or "economic," though "other possible kinds should not be overlooked." Id. "Risks and benefits of research may affect the individual subjects, the[ir] families, . . . and society at large (or special groups of subjects in society)." Id. Benefits include treatment and the "knowledge to be gained from the research." Id.

151. Id.

152. Breault, supra note 146, at 16.

153. Id. at 15 ("Everyone in research is responsible for human subject protection. Institutional Review Boards (IRB) are unique in that this is their sole reason for existence."). IRBs are composed of individuals representing varied institutional stakeholders, including both clinical investigators and others who explicitly must not be investigators in order to better represent research subjects' interests. Id. at 18.

154. Id. at 17–18.

155. Id.

demic studies will undergo some IRB scrutiny to ensure scientific validity and thus, beneficence.[156] This is necessary because "[e]very study has some risk," a study's benefits must outweigh its risks to be beneficent, and a study can have no benefit without scientific validity.[157]

Under the Common Rule, studies conducted by HIPAA-covered entities using only EHR data are exempt from IRB review because it is believed that the main risk is to the subjects' privacy.[158] Academic investigators using EHR data need only comply with HIPAA's privacy protections (namely, de-identification) to minimize this risk.[159] Nonetheless, it is commonly understood that a determination of exemption from IRB review may only be made by an IRB and that researchers cannot make this determination for themselves.[160] Moreover, publication in any reputable journal requires IRB review.[161] Thus, regardless of the law, an IRB will review academic proposals for compliance with the Belmont Principles.

BigMedTech firms will correctly argue that any research they conduct is exempt from IRB review because they do not receive any federal funding and thus are not regulated by the Common Rule.[162] But this lack of ethical supervision nonetheless violates the Belmont Principles that HIPAA and the Common Rule are meant to safeguard. The scale of data collected and

---

156. For an example of this type of review, see IRB Inquiry Form, Kaiser Permanente Ctr. for Health Rsch., https://research.kpchr.org/Portals/1/Documents/Forms-Templates/IRB_Inquiry_Form_NW.docx [https://perma.cc/ABV7-TAU6] (last visited Oct. 2, 2020) ("Section One—Please provide a brief summary (about half a page) of your proposed project, including the purpose and the methods being used[.]").

157. Breault, supra note 146, at 18 ("Therefore, the IRB reviews the basic scientific validity of the study, to determine if the benefits outweigh the risks.").

158. 45 C.F.R. § 46.104(d)(4)(iii) (2019).

159. Off. for Hum. Rsch. Prots., Attachment B—Recommendations on the Interpretation and Application of §_.104(d)(4) the "HIPAA Exemption", HHS, https://www.hhs.gov/ohrp/sachrp-committee/recommendations/attachment-b-december-12-2017/index.html [https://perma.cc/Q2P5-VZX4] [hereinafter 104(d)(4) Guidance] (last updated Dec. 15, 2017).

160. See, e.g., Does Your Study Require IRB Approval?, Rutgers Univ., https://orra.rutgers.edu/irb-review [https://perma.cc/BW53-NT92] (last visited May 26, 2020) ("Investigators cannot self-exempt . . . . Only the IRB can determine if your research is 'Not Human Subjects Research' (meaning it does not require formal IRB Approval)."); Frequently Asked Questions, U.C. Irvine Off. of Rsch., https://www.research.uci.edu/compliance/human-research-protections/researchers/irb-faqs.html [https://perma.cc/5H5Y-9TSE] (last visited May 26, 2020) (stating a similar requirement); IRB FAQs for Survey Researchers, Am. Ass'n for Pub. Op. Rsch., https://www.aapor.org/Standards-Ethics/Institutional-Review-Boards/IRB-FAQs-for-Survey-Researchers.aspx#question2 [https://perma.cc/8W5M-W69E] (last visited May 26, 2020) (same).

161. See, e.g., Protection of Research Participants, Int'l Comm. Med. J. Eds., http://www.icmje.org/recommendations/browse/roles-and-responsibilities/protection-of-research-participants.html [https://perma.cc/B2HD-D57U] (last visited May 14, 2020) (stating that virtually all peer-reviewed journals require that researchers submit their methods for approval by an independent ethics board or IRB); Zozus et al., supra note 20 (same).

162. See 45 C.F.R. § 46.101.

analyzed by BigMedTech dwarfs data collection by AMCs.[163] It is critical to remember that scale in big data analysis translates to powerful results, both in the statistical sense and in the sense that BigMedTech's studies may have rippling effects throughout society.[164] These studies are unpublished, so one can only speculate as to what conclusions they contain. But by way of example: An insurer could use EHR data to conclude that people who have high blood pressure tend to be overweight and take more time off from work. They could then sell this report to employers who could begin screening out overweight applicants. No IRB will have reviewed the validity of such a study to ensure that its benefits outweigh its risks. Where such a study draws conclusions about disadvantaged minorities, it also violates the principle of justice by unduly burdening one group of people with the consequences of its conclusions.[165]

2. *BigMedTech Is Not Subject to the Constraints of Academic Research and Its Research May Be Methodologically Flawed.* — The above discussion foreshadows perhaps the most fundamental problem with BigMedTech's EHR-based research: It is also not subject to any of the extralegal constraints imposed on academic investigators. While private studies' legal exemptions from full IRB review may be justified, there is no justification for exemption from any and all ethical or scientific scrutiny.

The peer review and publication processes impose strict, formal constraints on academic researchers that are absent in the for-profit setting. The International Committee of Medical Journal Editors, an association of major global medical journals, issues comprehensive guidelines for research ethics, which are prerequisites to publication.[166] Peer review—required under these rubrics—is "the [unbiased, independent, and] critical assessment of manuscripts . . . by experts."[167] Scientific hypotheses are generally only accepted by the academic community after publication in a

163. See supra sections I.A.2 and II.A (exploring how and why some BigMedTech firms have hundreds of millions of patient records to use for research while most AMCs have far fewer).

164. See supra notes 21–26, 93–94 and accompanying text (discussing the power of big data EHR analysis which, among other things, could help track the spread of viruses such as the one at the center of the COVID-19 pandemic).

165. The Tuskegee study's use of Black men is a clear example of a violation of the principle of justice. The Belmont Report, supra note 145; see Hoffman & Podgurski, supra note 54, at 107–08 (detailing instances where Ashkenazi Jews and Black Americans have been the subjects of research, the results of which could be used to stigmatize them and extending the potential for such risk to de-identified data).

166. See Int'l Comm. Med. J. Eds., Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals 1, 3, 7 (2019), http://www.icmje.org/icmje-recommendations.pdf [https://perma.cc/R544-C88Z] [hereinafter Recommendations for Publication] (requiring disclosure of financial conflicts of interest related to research, IRB review, protection of research participants, and compliance with various ethics codes); Protection of Research Participants, supra note 161.

167. Recommendations for Publication, supra note 166, at 5.

peer-reviewed journal.[168] Peer reviewers evaluate "research protocols[] [and] plans for statistical analysis"; they may also conduct independent data analysis or request such analysis from an independent biostatistician.[169] Where an article promotes bad science, it can quickly be debunked before it has pernicious effects on patient care.[170] Publication itself is another bulwark against bad science; an author's peers will evaluate scientific validity and experimental design as they read an article.[171] The lack of comparable review increases the risk that private-sector research will promulgate harmful science and poses a direct risk to social and communal wellbeing.

The culture of academic research also imposes constraints to which private-sector researchers are not subject. AMCs, responsible for the vast majority of published medical research, are different from community medical centers in that physicians not only treat patients but also conduct research.[172] Physician-scientists are formally trained in medical ethics and are sworn to uphold the Hippocratic Oath, which "embodies the philosophy . . . that human life is inviolable."[173] Every time a physician suggests a course of treatment to a patient, they must first decide if the benefits outweigh the risks, and they will bear the judgement of their patients, patients' families, and colleagues should they fail to do so adequately.[174] This informs how physician-scientists formulate research questions, design and conduct studies, and review other publications.[175] Clinical care

168. Jacalyn Kelly, Tara Sadeghieh & Khosrow Adeli, Peer Review in Scientific Publications: Benefits, Critiques, & a Survival Guide, 25 J. Int'l Fed'n Clinical Chemistry & Lab'y Med. 227, 229 (2014).

169. Recommendations for Publication, supra note 166, at 5–6.

170. See Adam Feldman, Peer Review: What Is It and Why Do We Do It?, Med. News Today (Mar. 29, 2019), https://www.medicalnewstoday.com/articles/281528 [https://perma.cc/545T-6TJ7].

171. See Recommendations for Publication, supra note 166, at 6 ("Some people believe that true scientific peer review begins only on the date a paper is published."); cf. Nat'l Rsch. Council of the Nat'l Acads., Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences 28 (2003) ("[A] risk associated with publishing is that other researchers will use information presented in a paper to invalidate or question the author's own findings, and publish conflicting results.").

172. What It Means to Be an Academic Medical Center, Penn Med., https://www.pennmedicine.org/about/benefits-of-an-academic-medical-center [https://perma.cc/88JA-ANRD] (last visited May 14, 2020); see also supra note 110 and accompanying text.

173. Stephen J. Genuis, Dismembering the Ethical Physician, Postgraduate Med. J., Apr. 2006, at 233, 233; Eve Glicksman, "What Do I Do?" Teaching Tomorrow's Doctors How to Navigate the Tough Ethical Questions Ahead, Ass'n Am. Med. Colls. (Sept. 27, 2016), https://www.aamc.org/news-insights/what-do-i-do-teaching-tomorrows-doctors-how-navigate-tough-ethical-questions-ahead [https://perma.cc/RY98-DJHE].

174. See Glicksman, supra note 173 (illustrating the importance of informed consent by recounting an instance where an attending physician did not correctly explain the risks and benefits of a particular treatment to a patient who subsequently died).

175. See Mukesh K. Jain, Tadataka Yamada & Robert Lefkowitz, Opinion, We Need More Doctors Who Are Scientists, N.Y. Times (Sept. 23, 2019), https://www.nytimes.com/

informed the development of penicillin,[176] statins, cancer and HIV treatments, and many more lifesaving therapies.[177]

The risk that unpublished for-profit medical research promulgates bad science is real. Data scientists working in for-profit settings have no ethical or clinical training, nor any professional constraints on their activities, despite the immense power of the data they analyze.[178] More to the point, in addition to owing no duty to the patients whose records they are studying, BigMedTech firms are bound to maximize shareholder value.[179] While profit maximization plays a crucial role in the development of thousands of lifesaving drugs and devices,[180] it could clash with the Belmont Principles. For example, a BigMedTech firm could alter its research methodology to reach the answer sought by its client to avoid losing future business. Repeated instances of such behavior would warp its conclusions and influence the healthcare industry to the detriment of patients and society.[181] Thus, its opacity and lack of review inherently renders BigMedTech's research ethically and scientifically problematic, regardless of whether any individual studies have been proven to be harmful in any other sense.

Moreover, for-profit research likely leaves important medical questions unasked and unanswered. Academic medical research often focuses on the development of new (and lucrative) therapeutic interventions (drugs, devices, and more).[182] But academic investigators also study "disease trends and risk factors, outcomes of treatment or public health interventions, functional abilities, patterns of care, and health care costs and

---

2019/09/23/opinion/doctor-scientist-medical-research.html (on file with the *Columbia Law Review*) ("Time and again, physician-scientists have changed the history of medicine by identifying a problem in the clinic and taking to the lab to address it.").

176. Id.

177. The Future of Medicine Depends on Physician-Scientists, Physician-Scientist Support Found., http://www.thepssf.org [https://perma.cc/87BQ-H779] (last visited May 26, 2020).

178. See Vin Vashishta, Why Don't Data Scientists Get Ethics Training?, Silicon Republic (Apr. 10, 2018), https://www.siliconrepublic.com/careers/data-scientists-ethics [https://perma.cc/2KVR-YLV9].

179. Lucian Arye Bebchuk, Federalism and the Corporation: The Desirable Limits on State Competition in Corporate Law, 105 Harv. L. Rev. 1435, 1492 (1992) ("State corporate law in this country, however, has traditionally taken the position that the managers' duty is to serve shareholders' interests—specifically, the maximization of long-run profits.").

180. Research!America, 2018 Annual Report 11 (2018), https://www.researchamerica.org/sites/default/files/Research%21America_AnnualReport18_onlineversion.pdf [https://perma.cc/Z28X-LTE2] (finding that the $121.8 billion cost of private investment in medical research and development is dwarfed by the return on these research and development costs in the form of valuable new drugs and treatments).

181. That is because without scientific validity, research results are unreliable and thus can produce no benefit to outweigh the risks they pose. See supra notes 156–157 and accompanying text.

182. See Inst. of Med., Comm. on Health Rsch. & the Privacy of Health Info., Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research 20–21 (Sharyl J. Nass, Laura A. Levit & Lawrence O. Gostin eds., 2009).

use."[183] These subjects may not always lead to profitable discoveries, but their study can nonetheless save lives. For example, the healthcare disparities between rural and urban areas is an important subject of medical research, but it is doubtful a private company would be able to market such a study.[184]

Collaboration between medicine's private and academic sectors is credited with directly leading to a five-year (or seven percent) increase in American life expectancy between 1970 and 2000.[185] In economic terms, it is estimated that these results are worth nearly $3 trillion a year.[186] While BigMedTech firms may have excellent data scientists at their disposal, big data techniques are just that: techniques.[187] In order for these tools to be put to their best use—replicating the twentieth century's leap in medical knowledge—they must be integrated into clinical research, but they cannot replace it.[188]

III. UNIFYING THE REGULATION OF EHR-BASED RESEARCH

Legal commentators agree that big data EHR-based clinical research is a "public good," the facilitation of which should be among any data-governance regime's primary goals.[189] But previous scholarship on EHR data regulations has focused on fixing the problem of "data fragmenta-

183. Id. at 112.

184. See id. at 20 ("Medical records research has documented that disparities and lack of access to care in inner cities and rural areas results in poorer health outcomes, and has demonstrated that specific preventive services (e.g., mammography) substantially reduce mortality and morbidity at reasonable costs.").

185. Leon E. Rosenberg, Exceptional Economic Returns on Investments in Medical Research, 177 Med. J. Austl. 368, 368 (2002).

186. Id. at 371.

187. Big data analysis, while promising, is not magic; it is a set of advanced computer science and predictive statistics techniques with its own biases, much like traditional statistical methods. See Daniel J. Grimm, The Dark Data Quandary, 68 Am. U. L. Rev. 761, 819–20 (2019) (discussing the "built-in constraint[s] that preclude[] Big Data-derived conclusions from deserving the gloss of fact-inclusive omnipotence they often receive").

188. See Harlan M. Krumholz, Big Data and New Knowledge in Medicine: The Thinking, Training, and Tools Needed for a Learning Health System, 33 Health Affs. 1163, 1163–70 (2014) (discussing big data's potential to revolutionize clinical medicine but assuming throughout that such research must be led by academic scientists).

189. See Jorge L. Contreras, The False Promise of Health Data Ownership, 94 N.Y.U. L. Rev. 624, 641 (2019) (arguing that propertization of EHR data could "impede socially valuable biomedical research"); Evans, Barbarians at the Gate, supra note 93, at 654 ("Twenty-first century science . . . needs large-scale, deeply descriptive, and inclusive data resources."); Mark A. Hall, Property, Privacy, and the Pursuit of Interconnected Electronic Medical Records, 95 Iowa L. Rev. 631, 640 (2010) (arguing that medical research can make productive use of EHR data); Hoffman & Podgurski, supra note 54, at 114–23 (arguing that individual consent rights distort research and make it prohibitively expensive); Pasquale, Grand Bargains, supra note 70, at 737 (arguing that secondary research on EHRs may eventually become more valuable than clinical trials once researchers have access to sufficient quantities of data); Rodwin, supra note 65, at 595–99 (arguing that public data ownership would lead to superior research outcomes).

tion."[190] For example, Professor Marc Rodwin acknowledges that firms traffic in EHR-based research, but goes on to describe the proliferation of "databases fractured among sub-populations[] [of patients]" and thus poorly suited to population-wide research.[191] But in fact, the above analysis demonstrates that massive EHR databases exist and that they owe their existence to the Privacy Regime.[192] Instead, the problem is that the Privacy Regime has left EHR data largely in private hands and outside of academic medicine, subverting foundational principles of sound research design and ethics.

The solution to this problem is relatively straightforward: Congress must pass a law to impose a basic form of IRB-like ethical supervision on private EHR-based research and provide academic researchers with access to EHR databases. Section III.A argues for using IRB review and public disclosure to impose transparency on BigMedTech firms' data collection and use. Next, section III.B argues the law should leverage the Privacy Regime's EHR aggregating mechanisms to give clinical investigators access to EHR data for research projects. Indeed, the EHR data market will be integral to this scheme to avoid needless duplication of work already accomplished by BigMedTech under the Privacy Regime. In exchange for the continued right to profit from EHR data, BigMedTech must submit to basic ethical supervision and provide academic researchers with access to their databases.

Some may believe that the existing market for de-identified medical data will continue maturing and eventually resolve the data distribution concerns section II.B identifies. But as Professor Frank Pasquale has argued, contemporary big data collection and analysis initially developed in the open but became secretive when transparency became a competitive disadvantage.[193] Indeed, BigMedTech's commitment to secrecy is evident: While these firms may be willing to publicize their commitment to using

---

190. See Evans, Barbarians at the Gate, supra note 93, at 667 ("In traditional healthcare and research environments, control of data remains fragmented among multiple data holders . . . . Harnessing data for public good requires transactions to bring the data together."); Hoffman & Podgurksi, supra note 54, at 128–30 (discussing various means of creating a centralized EHR database but not the existence of private EHR databases); Pasquale, Grand Bargains, supra note 70, at 683 ("By siloing data, health insurers and providers have impeded the types of large-scale analysis common in other industries."). CMS Administrator Seema Verma echoed that concern when proposing CMS's new interoperability rule. CMS Advances Interoperability & Patient Access to Health Data Through New Proposals, CMS Newsroom (Feb. 8, 2019), https://www.cms.gov/newsroom/fact-sheets/cms-advances-interoperability-patient-access-health-data-through-new-proposals [https://perma.cc/YQ8K-E4SP].

191. Rodwin, supra note 65, at 601.

192. See supra sections I.A.2, II.B.

193. See Frank Pasquale, The Black Box Society: The Secret Algorithms that Control Money and Information 193 (2015) ("[K]nowledgeable but unscrupulous individuals learned how to game exposed systems, and the profit advantage of informational exclusivity was too strong to resist. The less known about our algorithms . . . the better . . . . Transparency was replaced by ironclad secrecy, both real and legal. The matter of legitimation was tabled.").

EHR data for public health research,[194] what little is actually known about their data practices was revealed only thanks to Tanner and a few others' investigative reporting.[195] The solution presented below aims to permit BigMedTech firms to continue selling data analyses without sacrificing all competitive advantages. But ultimately, ethics in research requires transparency and competitive firms will not commit to transparency unless required to do so.

A. *Sunlight Is the Best Disinfectant: Transparency and Review Protects Patients*

Transparency is a well-established means of reforming data markets. Professor Neil M. Richards and Jonathan H. King argue that a theory of "big data ethics" must be built as much on transparency as on privacy to prevent "abuses of institutional power" and to encourage individuals to feel safe in sharing their data.[196] The European Union Data Protection Directive uses transparency to track malicious data usage and enable law enforcement—one cannot police what is invisible.[197] Moreover, imposing transparency on private enterprise is likely less cumbersome than other proposed reforms, such as providing users with the opportunity to opt out of data collection, which would likely inhibit data aggregation.[198]

Transparency in EHR data research would help regulators, academics, and the public monitor BigMedTech firms, mitigating the risk of potential data abuses.[199] This Note can only identify the legal mechanisms that enabled BigMedTech to amass troves of EHR data, and discuss uses of that data thanks to the work of investigative journalists such as Adam Tanner, and a few reporters at the *Wall Street Journal*.[200] But if BigMedTech firms were forced to disclose how they collect, use, and sell data, future scholars

194. See Tanner, supra note 27, at 143 (quoting Cerner's CEO arguing that research on de-identified medical record data by pharmaceutical companies is the "basis of advancement in medicine").

195. See supra notes 27–45 and accompanying text.

196. Neil M. Richards & Jonathan H. King, Big Data Ethics, 49 Wake Forest L. Rev. 393, 396 (2014).

197. Directive 95/46/EC, of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, 1995 O.J. (L 281) 31, 38.

198. See Omer Tene & Jules Polonetsky, Big Data for All: Privacy and User Control in the Age of Analytics, 11 Nw. J. Tech. & Intell. Prop. 239, 242 (2013) (discussing transparency's greater flexibility relative to individual consent requirements). Similarly, patients may prefer transparency rules over adding healthcare to the list of industries that permit "opting into or out of seldom read, much less understood corporate privacy [and data collection or use] policies." Id.

199. See supra section II.B (discussing how BigMedTech's unfettered analysis of EHR data subverts the Belmont Principles).

200. See supra notes 27–45 and accompanying text.

would be able to better analyze the possible harms and benefits of that research.[201]

Thankfully, imposing transparency on private medical research does not require a radically new regulatory regime, merely an extension of the existing IRB review process. First, as other scholars have suggested, the Common Rule's protections should cover all EHR-based research, not just federally funded studies.[202] Second, while the Common Rule's HIPAA exemption should still apply to EHR-based research, the Rule should explicitly specify that only IRBs can make exemption determinations.[203] This reform would not eliminate the Common Rule's efficient HIPAA exemption.[204] Rather, it would eliminate the regulatory distinction between research using ex ante de-identified data (research not currently subject to any Common Rule or HIPAA regulation) and research that starts with PHI (but is then de-identified to comply with HIPAA).[205] Any research proposal meeting the statutory definition of research would be reviewed by an IRB created and funded by the firm where the proposal originates.[206] Formally, the review would entail no more than confirming

---

201. See supra note 150 (discussing the principle of beneficence, which requires weighing a proposed research project's benefits against its harms).

202. Commentators have suggested extending the Common Rule to correct "policy drift" that has occurred as biomedical research funding has come increasingly from the private sector, instead of federal agencies. See e.g., Gabrielle Goldstein, Moving Beyond the Federal Funding Hook: Management-Based Regulation in Biomedical Research, 10 Drexel L. Rev. 127, 177–78 (2017).

203. Technically, there are two exemptions, one for HIPAA-covered identifiable EHR data and another for de-identified data (because the latter is not human subject research). See 45 C.F.R. §§ 46.102(e)(1)(ii), 46.104(d)(4) (2019). Some have suggested similar or more demanding extensions of the Common Rule. See Hoffman & Podgurski, supra note 54, at 91 ("[Studies] involving only de-identified data, which are currently exempt from scrutiny, [should] be reviewed and monitored by an ethics board with expertise in record-based research.").

204. See 104(d)(4) Guidance, supra note 159 ("[I]t seems appropriate from an ethical perspective, and less confusing and burdensome for researchers, if the activities already subject to HIPAA's rigorous requirements regarding research use of PHI not simultaneously be subject to the Common Rule's requirements regarding the use of identifiable private information.").

205. To be clear: In most instances, medical investigators identify the EHR data they wish to use for research and then approach their institution's IRB. See Khaled El Emam, Methods for the De-Identification of Electronic Health Records for Genomic Research, Genome Med., Apr. 2011, at 1, 1; see also supra notes 158–161 and accompanying text. That IRB will then allow a project to go forward only if all subjects have consented to the research or if data are de-identified before research begins. See El Emam, supra, at 1. This functionally means that most academic EHR-based research, like private research, is conducted on de-identified data—but unlike private research, it is subject to IRB oversight, despite there being no meaningful difference in data aggregation and use. See supra notes 100–120 and accompanying text.

206. See 45 C.F.R. § 164.501 (defining research as "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge"); Todd H. Wagner, Aman Bhandari, Gary L. Chadwick & Daniel K.

that the proposed research will use only de-identified, previously collected information (i.e., EHR data).[207] Submissions for review, however, would have to explain the question being investigated, even if the question itself is not being formally reviewed.[208] This requirement, even without more intensive review, encourages self-enforcement of the Belmont Principles.[209]

Finally, this law should require that BigMedTech firms periodically provide the National Institutes of Health (NIH) (or another agency) with reports describing their data holdings and summarizing all research conducted or authorized. This would provide some semblance of the post-publication peer-review process to which academic investigations are submitted.[210] Academics, government officials, and the public would have the chance to see how EHR data were used in research. It would help regulators better understand the nature and scope of private EHR research and decide if more stringent measures are appropriate. Second, to the extent harmful research or sales practices are not screened out by IRB review, public outcry may help cabin such excesses.[211] But the contents of this publication would have to be fine-tuned to protect against overdisclosure, which might risk diminishing a firm's competitive edge and thus overdiscouraging valuable research and data aggregation.[212]

B. *Providing Researchers with Access to the EHR Data Mine*

Providing researchers with access to BigMedTech's data stores presents a few distinct challenges. First, robust data collection must continue; otherwise there will be no data to use. For that to happen, BigMedTech

---

Nelson, The Cost of Operating Institutional Review Boards (IRBs), 78 Acad. Med. 638, 638 (2003) (describing IRB costs as costs to the "medical center" that houses the particular IRB).

207. Again, this is already standard practice in the medical research community. See supra notes 160–161 (listing institutions that require an IRB determine if a study is exempted from full IRB review).

208. See supra notes 160–161 and accompanying text.

209. See Hannah L. Baldwin, Note, Clearing the Air: How an Effective Transparency Policy Can Help the U.S. Meet Its Paris Agreement Promise, 35 J.L. & Com. 79, 95–96 (2016) ("The goal of a transparency policy as a regulatory mode is to promote self-regulation within regulated entities through the dissemination of data."); Russ Linden, Transparency Breeds Self-Correcting Behavior, Governing (Jan. 13, 2010), https://www.governing.com/columns/mgmt-insights/Transparency-Breeds-Self-Correcting-Behavior.html [https://perma.cc/6JL7-E3BK] (last updated Feb. 1, 2020) ("Operating in a transparent way offers enormous benefits to government agencies and to the public. It can improve operations, increase accountability and raise trust.").

210. See supra notes 166–171 and accompanying text (discussing the scrutiny to which academic papers are submitted after publication and its value in protecting the Belmont Principles).

211. See Linden, supra note 209 (discussing various examples where transparency revealed poor institutional practices, leading to public outcry, and subsequently to improved behavior in the regulated entities).

212. Cf. Richards & King, supra note 196, at 420–21 (discussing the "inherent[] . . . tension" between transparency and secrecy where secrecy may protect valuable trade secrets).

firms need continued incentives to aggregate data, which means letting them continue participating in the EHR data market. Second, investigators need an efficient mechanism for acquiring the right data for their proposals.

First, BigMedTech firms should be allowed to continue aggregating data and selling research products to pharmaceutical firms and other customers, subject to IRB review and reporting. While legal commentators agree that providing scientists with access to EHR data is a public good, most other proposals suggest various new regulatory regimes aimed at incentivizing aggregation.[213] But as discussed above, the EHR data market and the Privacy Regime create powerful incentives and means for data aggregation.[214] Existing proposals include private data ownership,[215] public data ownership (where all EHR data are managed by a federal agency),[216] and hybrid forms of data aggregation and rights management.[217] These proposals have varying degrees of merit, but they falter in replacing the efficient (albeit unappealingly secretive and profit-motivated) private system of data aggregation with more cumbersome schemes. EHR data management is a highly complex, technical, and expensive undertaking.[218] Building an entirely new command-and-control style data aggregation apparatus risks rendering data even less accessible to researchers than they are now.[219]

---

213. See supra note 189 and accompanying text (listing examples).

214. See supra sections I.A.2–II.A.

215. Professor Mark Hall recommended giving patients the power "to authorize access to and use of their medical information for financial rewards, . . . [which would enable society to capture] network benefits [and] . . . incentive[ize] [database] construction." Hall, supra note 189, at 660. Critics fear that Hall's proposal would inhibit forming "comprehensive databases . . . [by facilitating] data monopolies that will increase the price of data" and research. See Rodwin, supra note 65, at 589; see also Evans, Barbarians at the Gate, supra note 93, at 667 (noting that when individuals control their data, "[h]arnessing [that] data for public good requires [cumbersome] transactions to bring the data together"). Moreover, EHR data are poorly suited to propertization because they lack "exclusivity, infinite duration, divisibility, [and] alienability." Contreras, supra note 189, at 633.

216. Rodwin, supra note 65, at 615 (suggesting providers and other healthcare industry players report EHR data to "the []HHS or a[nother] public authority . . . as they do to third-party payers when seeking payment").

217. Professor Barbara Evans suggested the novel approach of creating "consumer-driven data commons" to aggregate "data for a group of participating volunteers who, thereafter, would employ processes of collective self-governance to make decisions about how the resulting data resources . . . can be used." Evans, Barbarians at the Gate, supra note 93, at 654.

218. See supra notes 125–132 and accompanying text (discussing the expense and technical challenges faced during the data aggregation process).

219. This is arguably the trap into which interoperability has fallen: fighting the economic dynamics of the healthcare industry rather than working with them. See supra notes 133–134 and accompanying text (discussing the reasons various stakeholders have to resist incentives to adopt or create interoperable EHR).

Thus, despite how morally appealing it might be to take direct control of EHR data,[220] the risk of destroying a resource worth potentially trillions of dollars in saved lives is simply too great.[221] Instead, data should remain in the hands of BigMedTech firms and a system of data access for academic research should be created separately. With dozens or hundreds of firms and thousands of investigators, there needs to be a central clearinghouse to manage this challenge.

One option is to let the NIH simply assign research projects to BigMedTech firms using the data disclosures previously proposed.[222] This solution might not be optimal for a few reasons. First, it may falter in the face of persistent informational asymmetries—the NIH may know something about a particular firm's data thanks to its disclosures, but inevitably it will know less than the firm itself. Second, it does not provide an incentive for firms to willingly cooperate with either government agencies or academic institutions.

A less blunt instrument might allow BigMedTech firms to bid on research "contracts." This would likely require added incentives, perhaps in the form of direct payments or tax credits. The firm presenting the most optimal combination of low bid price and high data quality for the research project would be awarded the contract. Over time, various private and academic institutions might form good working relationships (and approach the NIH together, with prepackaged deals) that could lead to improved EHR design, data aggregation practices, and efficiency of the data-distribution system.[223]

---

220. See Jaron Lanier & E. Glen Weyl, A Blueprint for a Better Digital Society, Harv. Bus. Rev. (Sept. 26, 2018), https://hbr.org/2018/09/a-blueprint-for-a-better-digital-society (on file with the *Columbia Law Review*) ("Most important, a market for data would restore dignity to data creators, who would become central to a dignified information economy."); see also Jaron Lanier, Jaron Lanier Fixes the Internet, N.Y. Times (Sept. 23, 2019), https://www.nytimes.com/interactive/2019/09/23/opinion/data-privacy-jaron-lanier.html (on file with the *Columbia Law Review*) ("You should have the moral rights to every bit of data that exists because you exist."); Stacey Vanek Smith, Should Social Media Companies Pay Us for Our Data?, NPR (Apr. 12, 2018), https://www.npr.org/2018/04/12/601759872/should-social-media-companies-pay-us-for-our-data?t=157720913909 [https://perma.cc/PNB5-ZXS8] ("I want to see companies compete and say to people, look, you shouldn't be taken advantage of. We will pay you a fair price for your data.").

221. See supra notes 185–186 and accompanying text (discussing the trillion dollars in value contributed by biomedical research in the last third of the twentieth century).

222. The NIH may be the best suited agency for such a task, given its current role in reviewing 50,000 competitive grants and awarding $41.7 billion annually. Budget, NIH, https://www.nih.gov/about-nih/what-we-do/budget [https://perma.cc/38VJ-L7N3] (last updated June 29, 2020). The NIH grant process includes a robust prefunding peer-review process to ensure that only high-quality science is funded. See Grants Process Overview, NIH, https://grants.nih.gov/grants/grants-process-overview.pdf [https://perma.cc/2JXY-WP9M] (last visited May 23, 2020). The same process should apply in this system.

223. See supra notes 185–186 and accompanying text (discussing the trillion dollars in value created by private–academic biomedical research partnerships in the last third of the twentieth century).

While there are certainly obstacles to the effective operation of such a scheme, an important one is the risk that firms become reticent to share data. Data sharing might harm a firm's competitive advantage, depressing the value of its data on the EHR data market to a point where data aggregation is no longer profitable.[224] Awarding direct payments or tax incentives may mitigate some profit loss. But whatever scheme is created for private–academic data sharing should also safeguard the EHR data market. The solution may be to provide researchers with access to opaque "data enclaves" to prevent the accidental dissemination of data (which risks depressing the value of a firm's data to private customers).[225] The challenge for the ultimate solution will thus be balancing access for research against the continued survival of the EHR data market.

CONCLUSION

Privacy regulations are being considered and implemented across a range of industries in an effort to crack down on morally dubious and dangerous uses of personal data.[226] Such efforts are admirable and important. But a world in which every industry is regulated by its own HIPAA is not better or more equitable merely because it is more private. As regulations erect barriers to data use, they create data pipelines, shape the flow of that data, and enrich some firms at the expense of society. This is particularly problematic in industries like healthcare where electronic data storage is essentially mandated by law, leaving individuals with no choice but to contribute sensitive information to an unknown firm for sale on the market.[227] Regulators and lawmakers must strive to use evolving data flows to correct negative externalities (in this case, the creation of an unregulated realm of medical research) and ensure that data are available to academic or public sector institutions for valuable research. Perhaps counterintuitively, such laws should avoid dismantling data markets. Rather, they should recognize the existence of such markets as proof of the data's value and leverage existing legal and economic dynamics to both protect individuals and encourage beneficial research.

---

224. See supra notes 192–194 and accompanying text (noting that the EHR data market, like other such markets, developed in secrecy because secrecy is a competitive advantage).

225. See supra note 38 and accompanying text (noting that Cerner sells access to such data enclaves that allow investigators to analyze the data without actually seeing it directly).

226. The California Consumer Privacy Act, for instance, grants Californians wide latitude in opting out of online data collection for a variety of activities. See Cal. Civ. Code §§ 1798.100–1798.120 (2020). For an example of a proposed privacy law covering all U.S. data collection, see Nuala O'Connor, Reforming the U.S. Approach to Data Protection and Privacy, Council on Foreign Rel. (Jan. 30, 2018), https://www.cfr.org/report/reforming-us-approach-data-protection [https://perma.cc/36FP-473U].

227. See supra notes 67–70 and accompanying text.