YALE LAW & POLICY REVIEW INTER ALIA

Derived Data: A novel privacy concern in the age of advanced biotechnology and genome sequencing

*Alda Yuan**

Cheap genetic sequencing, big data, and advanced biotechnology have the potential to revolutionize healthcare, but they also raise health data privacy concerns. They permit the emergence of derived data, which is unknown to the individual it describes and obtained through the analysis of existing data, both related and unrelated to healthcare. Derived data implicates the effectiveness of informed consent, the current method to protect patient privacy. Patients, research subjects, and consumers cannot reasonably consent to sharing, analysis, or use of data they do not know exists. To protect privacy rights while enabling progress in healthcare, regulations which now conceptualize data in silos must properly contend with 21st century data processing capabilities to link distant and seemingly unrelated data to form a more compete whole.

INTRODUCTION

It is, by now, trite to say that modern-day health-data regulations are inadequate for meeting the challenges of twenty-first-century medicine.[1]

* Alda Yuan received her J.D. from Yale Law School in 2018 and will be completing a fellowship at the Environmental Law and Policy Center in Chicago (but only after she passes the bar).

1. *See, e.g.*, Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1731-32 (2010) (discussing how re-identification should cause us to reconsider privacy law); Christina Scelsi, *Care and Feeding of Privacy Policies and Keeping the Big Data Monster at Bay: Legal Concerns in Healthcare in the Age of the Internet of Things*, 39 NOVA L. REV. 391, 407 (2015) (discussing the risk of data breaches for healthcare).

This observation is especially true in the privacy arena.[2] This Essay draws from existing literature to describe how advances in biotechnology, big data, and the new healthcare methods they enable give rise to an entirely novel privacy concern: derived data.

Derived data is information about an individual that is unknown to that individual, but which can be extrapolated from existing data. This information is not obvious, but instead requires advanced statistical analysis and the cross-referencing of different sources of data. Derived data may involve concrete health characteristics such as bone density as well as statistical information such as risk of disease. Since the patient is herself unaware of this information, she is also necessarily unaware when it has been revealed. This lack of knowledge raises a number of concerns. For one thing, it implicates the effectiveness of informed consent, as it means that the patient is not fully informed. In addition, while health and privacy regulations tend to recognize that average individuals cannot personally guard against all misuse of their data, even limited responsibility for data security may be unreasonable if the individual does not know what information is accessible. Misuse of health data and the difficulty of managing informed consent are themselves not new. However, derived data has the potential to make these problems more intractable. It adds an additional layer of uncertainty that makes unauthorized and unethical uses of private data harder to identify. Failure to be attentive to this uncertainty threatens the abundant positive potential of using the proliferation of data to better tailor treatments to the individual.

The phenomenon of derived data is emblematic of the privacy threats of twenty-first-century medicine, and the inability of existing regulations to grapple with its wide-ranging consequences reveals an important flaw: the siloed and decentralized nature of the regulations. This inadequacy is dangerous not just for individuals whose data may be put at risk. It also endangers the systems-level healthcare revolutions modern technology can enable, which have the potential to vastly improve healthcare outcomes.

Part I briefly describes existing health-data regulations and outlines some criticisms. Part II discusses the major technological advances facilitating the development of new models of healthcare. Part III discusses how these advances also create a novel threat to privacy against which

2. *See generally* Ira S. Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L. REV. 703, 708-09 (2016) (describing the failure of anonymization and stagnant nature of privacy law).

existing regulations do not protect. Finally, Part IV advocates for a new legal framework for the protection of health data.

I. STATE OF PRIVACY REGULATIONS

A. HIPAA Privacy Rule

The Privacy Rule[3] was promulgated by the Department of Health and Human Services to implement the Health Insurance Portability and Accountability Act (HIPAA), which Congress passed in 1996.[4] The regulation is fairly narrow, and some commentators have said it is misleading to call it a "privacy rule" as this gives the false impression that it adequately preserves patient privacy.[5] The Privacy Rule covers Protected Health Information (PHI), a subset of health information, such as demographic data, which can be used to uniquely identify an individual.[6] The Privacy Rule applies only when the information is collected by certain entities,[7] including healthcare providers, health plans, employers, and healthcare clearinghouses.[8] The basic structure of the Privacy Rule is that when covered entities are in possession of PHI, they have certain duties that prohibit data disclosures for some purposes, such as sale, without obtaining express authorization from the individual.[9]

Putting aside for a moment the numerous exemptions under which a covered entity may disclose PHI without authorization,[10] which are themselves concerning, there are two significant flaws with the above system. The first is simply that the Privacy Rule does not cover enough

3. 45 C.F.R. pt. 164 (2017).

4. Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936.

5. *See* Nicholas P. Terry, *Protecting Patient Privacy in the Age of Big Data*, 81 UMKC L. REV. 385, 386 (2012).

6. 45 C.F.R. § 160.103 (2017).

7. *See* Scelsi, *supra* note 1, at 421.

8. *See* Lara Cartwright-Smith et al., *Health Information Ownership: Legal Theories and Policy Implications*, 19 VAND. J. ENT. & TECH. L. 207, 228 (2016).

9. *See id.* at 229.

10. *See* Deven McGraw & Alice Leiter, *A Policy and Technology Framework for Using Clinical Data to Improve Quality*, 12 HOUS. J. HEALTH L. & POL'Y 137, 141 (2012).

entities.[11] It forbids healthcare entities such as hospitals and their business associates from disclosing PHI—but it does not forbid disclosure by educational institutions or commercial entities (as long as they are not performing duties for covered entities),[12] even though these organizations also routinely handle PHI.

Second, because only the initial act of disclosure is regulated, there is no real recourse for otherwise protecting or recovering the data once it is disclosed.[13] For instance, there is no cause of action against entities that use the data once it has been released. This inadequacy may be relevant where there is a data breach or where some of the advanced technological tools discussed in Part II[14] are used to piece together and derive health data. To illustrate, imagine two identical sets of health information that describe an individual. One was collected by the individual's doctor; the other was pieced together by combining information pulled from publicly available information. The first set is protected by HIPAA; the second is not.[15] Since HIPAA covers entities rather than the content of the data itself, it fails to accomplish the goal of protecting patient privacy.

B. Genetic Information Nondiscrimination Act

The Genetic Information Nondiscrimination Act (GINA) of 2008 states that genetic information is PHI and is therefore protected under HIPAA.[16] Additionally, it adds two specific prohibitions relevant here. First, health plans and insurers are not permitted to use genetic information to make decisions about coverage and are generally prohibited from even requesting genetic information.[17] Second, employers are not permitted to

11. *See* Lawrence O. Gostin & Sharyl Nass, *Reforming the HIPAA Privacy Rule: Safeguarding Privacy and Promoting Research*, 301 JAMA 1373 (2009).

12. *See* Janine S. Hiller, *Healthy Predictions? Questions for Data Analytics in Health Care*, 53 AM. BUS. L.J. 251, 283 (2016) (explaining the scope of covered entities).

13. *See* Joshua D.W. Collins, *Toothless HIPAA: Searching for a Private Right of Action To Remedy Privacy Rule Violations*, 60 VAND. L. REV. 199, 201-02 (2007).

14. *See* discussion *infra* Section II.A.

15. *See* Terry, *supra* note 5, at 386 (explaining that constructed proxy profiles provide an end run around HIPAA).

16. 42 U.S.C. § 1320d-9 (2012).

17. *See* Cartwright-Smith et al., *supra* note 8, at 232.

use genetic information in hiring and firing decisions or to purchase such information.[18] These measures deal with a narrow set of concerns. For instance, an insurance company cannot cancel the policy of a person who, based on her genetic profile, has a fifty percent chance of developing Huntington's disease.[19]

Arguably, however, GINA does not meet the risks of twenty-first-century healthcare head-on. First, the entities and types of discrimination covered by the Act are neither sufficiently tailored nor broad enough. For instance, GINA does not cover the use of genetic data in certain types of health-related insurance, including those for long-term care, disability, and life.[20] In addition, GINA prohibits discrimination only in employment and in obtaining healthcare.[21] Other areas of life are left untouched. It would seem that the Act does not cover key aspects of civil rights such as housing, which may open the door for the use of genetic information to justify discrimination.

The more fundamental flaw is that derived data is unlikely to serve as the basis for the type of discrimination GINA imagines. Those in possession of derived data will not necessarily be interested in keeping someone from exercising a right. Derived data might be used to discriminate in another sense of the word—by targeting people for special solicitation. Imagine a calcium supplement company that was able to identify and target individuals with a genetic propensity toward lower bone density. Such an activity would not come under the reach of GINA but might still be a violation of fundamental privacy rights.

II. ADVANCES IN HEALTHCARE TECHNOLOGY AND THE HEALTHCARE REVOLUTION

A. Technological Innovations

Modern technology has changed practically every realm of human life. Healthcare is no exception. The subsections below detail some of the key technological innovations that are changing the way healthcare research is

18. *See* Hiller, *supra* note 12, at 287.

19. *See* Andrea Aiken, Note, *Contradiction in Terms: Genetic Nondiscrimination and Long-Term Care Insurance*, 53 U. LOUISVILLE L. REV. 597, 598 (2015).

20. *See* Sejin Ahn, *Whose Genome is it Anyway?: Re-identification and Privacy Protection in Public and Participatory Genomics*, 52 SAN DIEGO L. REV. 751, 777 (2015).

21. *See* Louise Slaughter, *Genetic Information Non-Discrimination Act*, 50 HARV. J. ON LEGIS. 41, 51 (2013).

done and the way treatment is delivered. Together, big data, cheap electronics, rapid gene sequencing, and electronic health records have the potential to revolutionize healthcare.

i. Big Data

There are two interrelated definitions for big data. According to the first and most basic definition, big data is simply an extremely large data set, of such a size and complexity that computers and programs of an earlier era would have been unable to handle and make appropriate use of it.[22] The second definition is data which is amenable to data mining and predictive analytics;[23] this definition incorporates the understanding that big data is not only about an increase in volume. The increase in computing power has been accompanied by a sophistication in statistical tools and learning algorithms capable of identifying trends that would not otherwise be obvious.

ii. Small, Cheap Electronics

The rapid spread of cheap microchips enables computers to be embedded into household objects as well as into portable monitors.[24] All of these devices can passively collect clinically relevant information outside of a formal clinical setting.[25] The data from these devices, which accommodate a variety of sensors, can provide a full accounting of an individual's day. This technology can be incredibly beneficial to healthcare, such as in the case of a patient who suffers from seizures preceded by certain detectable changes in vital signals. However, the potential for abuse and misuse is also great. Data breaches or back doors into the devices, which are designed to allow access to data, along with the ability to collect health data in volume and in circumstances far beyond clinical and research settings, will present new privacy challenges.

22. Wullianallur Raghupathi & Viju Raghupathi, *Big Data Analytics in Healthcare: Promise and Potential*, HEALTH INFO. SCI. & SYSTEMS, Feb. 7, 2014, at 1, 1.

23. *Id.*, at 2.

24. *See* Yasser Khan et al., *Monitoring of Vital Signs with Flexible and Wearable Medical Devices*, 28 ADVANCED MATERIALS 4373, 4387 (2016).

25. *See id.* at 4381.

iii. Fast, Cheap Gene Sequencing

In 2016, sequencing a genome cost roughly 0.0014% of what it did in 2001.[26] In large part, this shift can be attributed to the Human Genome Project, an international scientific effort that was launched formally in the 1990s and completed in 2003.[27] Its mission was to obtain a complete sequence of human DNA on a budget of three billion dollars.[28] As of 2016, the sequencing of a full human genome cost only one thousand dollars, and the price is predicted to continue to drop.[29] Along with advancements in biotechnology at large, genome sequencing has fueled immense excitement about targeted treatments such as gene therapy, which, by altering the genetic mutations that result in disease, may one day help to eradicate genetic illnesses.

iv. Electronic Health Records

The push to move all patients over to electronic health records (EHRs) received a boost with the passage of the 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act.[30] Under the HITECH Act, the federal government invested twenty-seven billion dollars to help incentivize hospitals and other health providers to switch to EHRs.[31] While still primitive in many respects,[32] in the aggregate, EHRs

26. *See* Tom Ulrich, *Opinionome: Can DNA Sequencing Get Any Faster and Cheaper?*, BROADMINDED BLOG (Sept. 13, 2016), http://www.broadinstitute.org/blog/opinionome-can-dna-sequencing-get-any-faster-and-cheaper [http://perma.cc/QFJ2-Z7J5].

27. *See* Leeroy Hood & Lee Rowen, *The Human Genome Project: Big Science Transforms Biology and Medicine*, GENOME MED., Sept. 2013, at 1, 2.

28. *See id.*

29. *See The Cost of Sequencing a Human Genome*, NAT'L HUM. GENOME RES. INST. (July 6, 2016), http://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome [http://perma.cc/5GRV-6476].

30. Health Information Technology for Economic and Clinical Health Act, Pub. L. No. 111-5, 123 Stat. 226, 467 (2009).

31. *See* Brian Schilling, *The Federal Government Has Put Billions into Promoting Electronic Health Record Use: How Is It Going?*, COMMONWEALTH FUND (June/July 2011), http://www.commonwealthfund.org/publications/newsletters/quality-matters/2011/june-july-2011/in-focus [http://perma.cc/4HZM-JSBS].

have real potential to improve healthcare outcomes because they can be analyzed using big data methodologies.[33] This Essay argues that such techniques can threaten privacy, but also that precisely because big data is so powerful, it can contribute to both generalizable knowledge and more targeted remedies for the individual. Along with the wealth of data collected from other sources, EHRs can form the pool of data to help catalyze a true learning healthcare system where information derived from clinical care is funneled into research-style analysis to provide a check on earlier research findings, better accommodate individuals whose genetic makeup calls for different treatments, and provide useful information in developing new treatments.[34]

B. Healthcare-System-Wide Innovations: Precision Medicine

Changes in healthcare technology can help to enable evidence-based, detail-oriented, and more accurate healthcare for each individual.[35] Portable health devices are a concrete example. Imagine a patient with a heart-related issue who wants to be able to provide her physician with a full, accurate record of her heart rate and physical activity. Rather than relying on the patient's own reports of her elevated heart rate, which may be distorted by memory, a doctor using a portable health device can readily identify heart-rate changes as well as the activity levels to which they correspond.

However, the more significant changes that technological advances enable are structural. They have the potential to alter how healthcare is delivered, how research is performed, and how the system learns. These changes are sorely needed, as our current model does not properly make use of modern technology to maximize health outcomes.[36] Researchers,

32. *See generally* A. Begoyan, *An Overview of Interoperability Standards for Electronic Health Records*, INTEGRATED DESIGN & PROCESS TECH., June 2007, at 1 (discussing how the lack of standardization in EHRs is an obstacle to the sharing of medical data and how advances in medical science "demand further changes in existing EHR standards").

33. *See* Raghupathi & Raghupathi, *supra* note 22, at 5.

34. *See* Samuel J. Aronson & Heidi L. Rehm, *Building the Foundation for Genomics in Precision Medicine*, 526 NATURE 336, 338 (2015).

35. *See id.* at 336.

36. *See generally* E.R. Hsu et al., *Cancer Moonshot Data and Technology Team: Enabling a National Learning Healthcare System for Cancer To Unleash the Power of Data*, 101 CLINICAL PHARMACOLOGY & THERAPEUTICS 613 (2017); Burke

both public and private, perform randomized control trials (RCTs) to test new therapies, new drugs, and new regimens for delivering those treatments.[37] This research is published in medical journals, which are rarely read by practitioners.[38] Only a small proportion of the research results published in scientific journals are verified by replication.[39] Nevertheless, new information developed this way is incorporated sporadically into the standard of care,[40] which physicians apply to patients whom they only know by charts and see only once or twice a year.

The most obvious problem with this structure is that there simply is not enough information to verify the results of the RCTs. One study found that over half of the most-cited papers published in major medical journals were at least partially contradicted by subsequent research, could not be replicated, or simply were not challenged by replication attempts.[41] While

---

W. Mamlin & William M. Tierney, *The Promise of Information and Communication Technology in Healthcare: Extracting Value from the Chaos*, 351 AM. J. MED. SCI. 59 (2016); Rohini A. Patil & Anant D. Patil, *Use of Information Technology in Healthcare Sector for Improving Outcomes*, 3 INT'L J. BASIC & CLINICAL PHARMACOLOGY 269 (2014).

37. *See* Ted J. Kaptchuk, *The Double-blind, Randomized, Placebo-controlled Trial: Gold Standard or Golden Calf?*, 54 J. CLINICAL EPIDEMIOLOGY 541 (2001).

38. See MICHAEL MILLENSON, DEMANDING MEDICAL EXCELLENCE: DOCTORS AND ACCOUNTABILITY IN THE INFORMATION AGE (1997); Moyez Jiwa, *Doctors and Medical Science*, 5 AUSTRALASIAN MED. J. 462 (2012); Richard Smith, *The Trouble with Medical Journals*, 99 J. ROYAL SOC'Y MED. 115 (2006).

39. *E.g.*, Joel Achenbach, *Many Scientific Studies Can't Be Replicated. That's a Problem*, WASH. POST (Aug. 27, 2015), http://www.washingtonpost.com/news/speaking-of-science/wp/2015/08/27/trouble-in-science-massive-effort-to-reproduce-100-experimental-results-succeeds-only-36-times [http://perma.cc/2CSP-AU66] (discussing an experiment in which researchers "attempted to reproduce the results of 100 experiments that had been published in three prestigious psychology journals," but succeeded only thirty-nine times).

40. *E.g.*, Laura E. Bothwell et al., *Assessing the Gold Standard – Lessons from the History of RCTs*, 374 NEW ENG. J. MED. 2175, 2177 (2016) (offering examples where information from RCTs is not always incorporated into the standard of care).

41. John Ioannidis, *Contradicted and Initially Stronger Effects in Highly Cited Clinical Research*, 294 JAMA 218, 218 (2005) (noting that sixteen percent of "highly cited original clinical research studies" were

at least some of these shortcomings might be due to capture and profit-motivated exaggeration, the fact of the matter is that incentives are not aligned for most research scientists to use their time and resources to replicate studies.[42] Using clinical data to do secondary research to verify the results of RCTs can provide much-needed verification.[43] Even if this verification occurs on a large scale, however, it will not cure all the flaws of the health system. RCTs, even when verified, can only reveal the response of the average patient due to the averaging that happens over large sample sizes.[44] Yet the "average patient" is merely a statistical creature. Few people will have the underlying health characteristics and response to treatment regimens that averaging assumes, so patients in real clinical situations are unlikely to react to treatment precisely like an average patient.[45] This problem is exacerbated for groups, such as minorities or pregnant women, that go underrepresented in clinical trials,[46] since their unique characteristics and responses to treatment often are not incorporated into the average.

The innovations of the past couple of decades can provide the basis for a new kind of healthcare system, one that caters to individual needs through precision medicine, but also, on a structural level, learns as much from clinical care as it does from research. This is possible if we, on the

---

contradicted by subsequent studies, sixteen percent had found effects that were stronger than those of subsequent studies, and twenty-four percent "remained largely unchallenged").

42. *Replication Studies: Improving Reproducibility in the Empirical Sciences*, ROYAL NETH. ACAD. ARTS & SCI. 43 (2018), http://www.knaw.nl/shared/resources/actueel/publicaties/pdf/20180115-replication-studies-web [http://perma.cc/D4SB-27DC].

43. *See generally* McGraw & Leiter, *supra* note 10 (explaining how clinical data can be used for this purpose).

44. PETER W. HUBER, THE CURE IN THE CODE: HOW 20TH CENTURY LAW IS UNDERMINING 21ST CENTURY MEDICINE 151 (2013).

45. *See* Richard L. Kravitz et al., *Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages*, 82 MILBANK Q. 661, 662 (2004).

46. *See generally* Meghan E. McGarry & Susanna A. McColley, *Minorities are Underrepresented in Clinical Trials of Pharmaceutical Agents for Cystic Fibrosis*, 13 ANNALS AM. THORACIC SOC'Y 1721 (2016); Alannah L. Phelan et al., *Exclusion of Women of Childbearing Potential in Clinical Trials of Type 2 Diabetes Medications: A Review of Protocol-Based Barriers to Enrollment*, 39 DIABETES CARE 1004 (2016).

one hand, apply big-data methodologies to the vast amounts of data that can be scraped from EHRs and portable medical devices and, on the other hand, make use of N-of-1 studies.[47]

N-of-1 studies take a single individual as their subject to find out which drug regimen, wellness tactic, or other healthcare measure works best for that person.[48] Thus, these studies are often better for the individual than RCTs, which don't address individual conditions but rather the most statistically common version of the condition in question. Moreover, the results of N-of-1 studies can still be generalized and extrapolated to apply to other individuals with relevant similarities. Big-data analytics can help reveal which similarities matter in terms of treatment efficacy and which do not. In combination, big data and N-of-1 research can contribute to precision medicine, a scheme in which medicine is tailored to the individual rather than to some ill-defined average patient.[49]

IV. DERIVED DATA: A NOVEL PRIVACY THREAT

If the technological advances discussed above fulfill even a fraction of their potential, the result will be a boon to healthcare outcomes. While we should not lose sight of these benefits, these advances also implicate privacy concerns. In particular, the breadth of health data and data with health implications being collected, coupled with inadequate regulation and poor security protocols, means that the modern age exposes individuals to a novel privacy threat—the risk that outside actors will act to produce derived data.

Mechanically, the creation of derived data involves using statistics and information processing to link together data from different contexts to make inferences and sophisticated guesses about individuals. The concept of derived data is similar to proxy health data. Third parties may use "big data [to] produce basically unprotected patient-level data that will serve

47. Nicholas J. Schork, *Personalized Medicine: Time for One-Person Trials*, NATURE (Apr. 19, 2015), http://www.nature.com/news/personalized-medicine-time-for-one-person-trials-1.17411 [http://perma.cc/L7WF-SJ3F].

48. *See* Naihua Duan et al., *Single-Patient (n-of-1) Trials: A Pragmatic Clinical Decision Methodology for Patient-Centered Comparative Effectiveness Research*, 66 J. CLINICAL EPIDEMIOLOGY S21, S22 (2013).

49. *See* Aronson & Rehm, *supra* note 34, at 336-37.

as an effective proxy for HIPAA-protected data."[50] Derived data is also similar to the idea of emergent medical data, which is produced through the analysis of data that has no apparent health content, such as Facebook posts or the text of emails.[51] Thus, health-related data emerges out of non-medical data.

Derived data is distinct from proxy data and emergent data because it centers on information whose content is unknown to the target individual. The concept of derived data is useful to policymakers and others concerned about data privacy because it helps to underline a major distinction between privacy risks old and new. The information involved in modern privacy risks might itself be unknown to the individual it describes. This difference raises a number of new issues and helps to further the argument that existing healthcare-data regulations require reorientation if they are to stay relevant and fulfill their protective function.

To understand how this privacy threat arises out of modern healthcare technology, imagine an individual who purchases a commercial genetic-testing kit to determine her ancestry and sends in a sample of her DNA. The consent forms she signs allow for her information to be "de-identified" and sold to a data broker. This "de-identification" process is legally effective in the sense that the process removes the pieces of information understood to be personally identifying under the law. Unfortunately, given modern data-processing techniques, data can be re-identified, once again allowing unique identification as well as linking data from multiple sources. Therefore, that same data broker can scrape publicly available data on the individual's shopping habits, the frequency of her hospital visits, and her prescriptions. Meanwhile, the company that hosts data for her portable fitness device sells her data pursuant to the terms and conditions she did not read. Combining this data can reveal new information. Perhaps her DNA makeup reveals a genetic illness of which she is unaware. Or perhaps a combination of family history, genetic makeup, and lifestyle choices give her a ninety percent chance of developing heart disease by the age of sixty. All of these conclusions are, of course, unknown to the individual involved.

50. Nicholas P. Terry, *Big Data Proxies and Health Privacy Exceptionalism*, 24 HEALTH MATRIX 65, 87 (2014).

51. Mason Marks, *Emergent Medical Data*, BILL HEALTH (Oct. 11, 2017), http://blogs.harvard.edu/billofhealth/2017/10/11/emergent-medical-data/ [http://perma.cc/QJU4-V47Y].

This hypothetical scenario implicates at least two major ethical problems. First, there is a consent problem, as no one can reasonably consent, either in research settings or through terms of use, to the sharing and use of data about which she knew nothing. Second, if individuals are unaware of the information being transmitted and aggregated, it will be harder for them to know when it is used to discriminate against them or to manipulate them into taking action, such as making purchases.

All the data usages in the scenario above are currently legal. The data that can be used as grist for derived data is everywhere, falling under different legal regimes and levels of security. Health information is not secure, and because of its high value,[52] it presents an attractive target to hackers and unethical data brokers. As state requirements vary widely, there is no unified law governing notification of victims in the event of a data breach.[53] Nor does the law require parties who obtain the data to refrain from selling it or using it in a number of discriminatory and unethical ways. None of the healthcare regulations in Section II would protect all the data implicated here. Our healthcare regulations are too focused on the technical mechanisms of where the information is coming from and who may disclose it. Instead, they should focus on the value of the data, the potential for abuse, or the context in which the data is used. Of course, it is no solution to attempt to lock down all data sharing—that would defeat the potential benefits offered by new technology described above.

V. SUBSTANTIVE HEALTHCARE-DATA PRIVACY RIGHTS

The threat that derived data poses to privacy makes it clear that modern healthcare-data privacy regulations need better to take into account connections between data collected in different places and under different circumstances. HIPAA, GINA, and other statutes and regulations contemplate health data in separate contexts. The harm against which they guard is having the data in one specific bucket revealed. But that is a privacy risk of the twentieth century. In the twenty-first century, the

52. Caroline Humer & Jim Finkle, *Your Medical Record is Worth More to Hackers than Your Credit Card*, REUTERS (Sept. 24, 2014), http://www.reuters.com/article/us-cybersecurity-hospitals/your-medical-record-is-worth-more-to-hackers-than-your-credit-card-idUSKCN0HJ21I20140924 [http://perma.cc/2B34-D4AK].

53. *See* Brandon Faulkner, *Hacking into Data Breach Notification*, 59 FLA. L. REV. 1097, 1104 (2007).

potential harm is having these buckets of data linked together to produce derived data used to categorize and even manipulate people in ways they are not even aware is possible. This is a systems-level privacy threat, and systems-level thinking is needed to counter it. In the modern era, the production of data is like an ecosystem. There are links between disparate pieces, nonobvious on their face but nevertheless important. Tapping into these links may enable healthcare to advance in leaps and bounds. However, for privacy in health data to have real meaning and substance, healthcare regulations need to recognize and accommodate the risks that derived data create.

In particular, the various healthcare-information privacy regimes need to be brought together. For instance, the Common Rule, which governs research on human subjects, also contains some privacy protections even if privacy is not the main target of the regulation. Individuals who participate in human-subjects research therefore fall under the overlapping protection of explicit privacy rules such as HIPAA and the Common Rule. While the recent update to the Common Rule contains an attempt at harmonization,[54] these piecemeal attempts to make the regimes fit together cannot keep up with technological advances. Rather, Congress and regulatory agencies should construct a whole new framework to ensure that healthcare-data privacy rights are substantive rather than procedural. This does not necessarily mean, as some have suggested,[55] that a property regime should be grafted into the healthcare-data context. Or at least, it does not mean that bundles of rights appropriate in other property contexts should apply wholesale without consideration of the particular balance of the competing public interests in improving healthcare and protecting individuals' privacy.[56]

To better guard against the various privacy threats posed by modern technology, some scholars have suggested constructing a new legal framework for health-policy data from the ground up using Fair Information Practice Principles (FIPPs).[57] The Federal Trade Commission developed these principles and the guidelines attached to them as recommendations regarding fair information practice in the electronic

54. Data already protected by HIPAA is exempt from certain Common Rule requirements. *See* Barbara E. Bierer et al., *Revised 'Common Rule' Shapes Protections for Research Participants*, 36 HEALTH AFF. 784, 787 (2017).

55. *See generally* Jorge L. Contreras, *Genetic Property*, 105 GEO. L.J. 1 (2016).

56. *See* Cartwright-Smith et al., *supra* note 8.

57. *See* McGraw & Leiter, *supra* note 10, at 157-60.

marketplace.[58] The five core FIPPs are: "(1) Notice/Awareness; (2) Choice/Consent; (3) Access/Participation; (4) Integrity/Security; and (5) Enforcement/Redress."[59]

Application of these principles to health data would involve committing to openness and transparency, specifying the purposes for which personal data is collected, allowing collection of data through lawful means, limiting use to specified purposes, guaranteeing individual access to personal information, ensuring accurate and complete data, carefully safeguarding personal data, holding entities accountable, and developing remedies to address violations.[60] Further study and engagement with the public is necessary to flesh out how these should work in practice. Unlike existing legal regimes like HIPAA and GINA, which are siloed from each other and attempt to address only discrete problems, the FIPPs offer a more comprehensive theory of the privacy values that might be threatened in the modern age. Thus, the FIPPs can serve as a baseline upon which to build a unified legal framework with substantive rights that preserves patient privacy while enabling the sorts of data analysis upon which a learning healthcare system should be built.

Such a framework would help to close the legal loopholes that permit violations of fundamental privacy and would allow policy makers to approach the health-data ecosystem systematically. Privacy tradeoffs can then be made in practical, pragmatic ways designed to promote the use of big data for improving health outcomes, not for manipulation through the use of derived data.

VI. CONCLUSION

Derived data's privacy problem reveals not only that current regulations are inadequate, but also that the proposals put forward to address the various problems of re-identification are incomplete. A common solution proposed to meet new privacy threats, of which derived

58. *See* Martha K. Landesberg et al., *Privacy Online: A Report to Congress*, FED. TRADE COMMISSION ii (June 1998), https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-report-congress/priv-23a.pdf [http://perma.cc/8R8L-J3XT].

59. *Id.* at 7.

60. *See Common Framework for Networked Personal Health Information: Overview and Principles*, MARKLE 4-5 (June 2008), http://www.markle.org/sites/default/files/Overview.pdf [http://perma.cc/V9VW-8W5V].

data is an illustrative example, is better and broader informed consent.[61] Yet informed consent cannot cover those uses of which an individual is herself unaware. A possible rejoinder here is that individuals donate health data all the time without knowing what it contains. For example, the average patient knows little about genetics. Moreover, individuals sign over data describing test results, frequently without truly understanding what it reveals. However, in most of these contexts, patients assume and are told that their data will be de-identified. But genetic information and the data to which it can be linked can be re-identified. Studies have demonstrated this is already possible using publicly available information.[62] By their very nature, derived data will be associated with an individual and can be used to understand aspects of the individual that he does not understand about himself. From a privacy standpoint, this outcome is quite different from donating data when there is no risk of association with the donor.

The risks that derived data pose to privacy need to be addressed, not simply for the good of individual privacy rights, but also to ensure the promise of twenty-first-century medicine. In the modern era, we possess the technology to take advantage of increasingly large pools of data that better reflect the complexity of healthcare as well as the differences between individuals. Technological innovation is no barrier to the donation and use of data that can assist the identification of more effective cures. Privacy risks should not be either.

61. *See* Christine Grady et al., *Broad Consent for Research with Biological Samples: Workshop Conclusions*, 15 AM. J. BIOETHICS 34, 35 (2015); Ryan Spellecy, *Facilitating Autonomy with Broad Consent*, 15 AM. J. BIOETHICS 43 (2015).

62. *See* Erika Check Hayden, *Privacy Loophole Found in Genetic Databases*, NATURE (Jan. 17, 2013), http://www.nature.com/news/privacy-loophole-found-in-genetic-databases-1.12237 [http://perma.cc/YXJ4-J8J9].