On Medical Data Privacy

A Sociotechnical Research Paper
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

T Read Schlomer

May 6, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

T Read Schlomer

Sociotechnical advisor: Peter Norton, Department of Engineering and Society

On Medical Data Privacy

The healthcare industry accounts for about 17 percent of U.S. GDP (CDC, n.d.). In the past twenty years, new software has been designed to store and process the vast amount of data produced by patient encounters. This data can aid research that saves lives or help allocate health care resources efficiently. It could be used to embarrass, discriminate or impose costs based on health status. How are privacy advocates, researchers, governments, data brokers and their customers competing to draw the line between legitimate uses of patient data and invasion of privacy? Technology and industry developments have progressed to the point that existing laws regarding consumer data privacy and more specifically patient data privacy are being called into question. Data brokers would rather their consumer data acquisition activities remain obscure to consumers but privacy advocates, journalists and researchers are shedding light on these activities through varying measures.

Review of Research

Thorpe and Gray (2015) provides a good overview of the patchwork of existing laws regarding privacy of patient data in the United States. It enumerates the exceptions permitted by law that allows disclosure of health information without consent and cites specific examples of activities considered beneficial. For example, the quality improvement exception facilitated the creation of Minnesota's Reducing Avoidable Readmissions Effectively(RARE) project. It "uses hospital claims data to flag potentially preventable readmissions." and has "prevented nearly 8000 readmissions." The article also discusses projects that use data that isn't subject to health information privacy laws. IBM and pharmaceutical firm UCB collaborated in 2012 to improve epilepsy care. The project used de-identifed data, which isn't subject to privacy law, from 1.5 million epilepsy patients to develop

predictive analytics for physicians to use to make treatment recommendations to their patients. The writers make a sound case for the benefits of big data in healthcare but they don't discuss the potential for its misuse and the effect it could have on patients in their role as consumers or citizens.

Tsuji (2020) investigates the impact to healthcare and privacy of the Medical Big Data Protection Act in Japan. It discusses how the law defines de-identified data, how it is handled under the new Act and compares this arrangement to its analogues in the United States. The article describes the Act as building on legislation established by the Act on the Protection of Personal Information (APPI) established in 2003 and its revision in 2015. This gives a sense of privacy law progression as evolving in response to technology. The interests of the Japanese government are briefly mentioned and the effects on medical and business sectors are stated but the broader picture of how all of these participants act to affect the use of big data in medicine is not addressed. This context is important to understand the social forces that make policies that impact lives. Availed of this context, the citizen of a democracy may exercise their right to express their interests in the matter.

Koontz (2015) examines the state of privacy protections around health information in America. The author touches on many of the participants in the matter. She describes the basics of established law in a review of the Health and Accountability Act of 1996 (HIPAA),the Privacy Rule and the Security Rule and how they allow for the sharing of health information between entities. Technological adoption, spurred by government incentives, led to widespread adoption of the Electronic Health Record (EHR) which changed the way clinicians document and share medical data about their patients. The spreading of health information beyond healthcare systems to other industries is addressed by Federal Trade Commission(FTC) reports. Apple is briefly mentioned as taking action by changing its licensing agreement with developers to "prohibit the sharing of information gathered through its health and fitness app." Referenced polls and surveys show public

concern for privacy amidst all these new means of processing their medical data. Even suggested practices for healthcare professionals to protect the privacy of those they treat are offered. My intention is to further expand on this work by introducing some of the major participants shaping data privacy policy in general and medical data privacy specifically.

The State of Data Privacy

Data Brokers collect information about individuals, package all of it together and sell it to other companies. Credit card purchases, web searches, social media posts, public records and more are being aggregated and sold to health insurance companies(Allen, 2018). This data can be combined with de-identified patient data to help these companies assess risk for covering the population in a specific area. HIPAA allows for patient data to be shared with entities who aren't directly caring for the patient or billing for services if it is de-identified. The Department of Health and Human Services hosts guidance on how to meet the requirements of de-identification on its website(HHS, n.d.). Generally, these companies state that they use the data to either help their members lead healthier lives (Loeb, 2018) or that the data isn't used to price individual health plans(Allen, 2018).

De-identified patient data is also used in the development of new drugs and therapy for diseases(Singh et al., 2018). Data broker IQVIA promotes its products that help researchers complete clinical trials(IQVIA, 2021). IBM Watson Health explains it uses de-identified data and artificial intelligence to supplement knowledge from clinical studies to find therapies for diseases like COVID-19 while protecting the privacy of patient data(Landis-McGrath & Ballard-Stuart, 2020). But how anonymous is de-identified patient data?

Researchers have reidentified datasets that have had unique identifiers removed. Sweeney (2000) describes a method for cross referencing medical data with voter registration data to identify a portion of the population in a given area. In 2005, in testimony to the Department of Homeland Security, Latanya Sweeney PhD, offered this statement: "… in 1997 I was able to show how medical information that had all explicit identifiers, such as name, address and Social Security number removed could be reidentified using publicly available population registers (e.g., a voter list). In this particular example, I was able to show how the medical record of William Weld, the governor of Massachusetts of the time could be re-identified using only his date of birth, gender and ZIP. In fact, 87% of the population of the United States is uniquely identified by date of birth (e.g., month, day and year), gender, and their 5-digit ZIP codes." (Sweeney, 2005). Culnane et al. (2017) conducted a study that details a method to re-identify patients based on demographic information like date of birth, known medical procedures, gender and state. The number of confident re-identifications are relatively small around 3 or 4 but the study was performed with 10% of a dataset of longitudinal billing records, both medical and pharmaceutical. Dr. Rubinstein, who worked with Culnane and Teague in the aforementioned study, offered this statement: "We can improve confidence by cross-referencing with a second dataset of population-wide billing frequencies. We can also examine uniqueness according to the characteristics of commercial datasets we know of, such as bank billing data."(Bennet, 2017). Data brokers have access to many datasets that fit this criteria. Other studies show how reidentification is possible with consumer data(Naranyan & Shmatikov, 2008)(de Montjoye et al., 2015) and location data(Atockar, 2014)(de Montjoye et al., 2013). Researcher de Montjoye describes the work he published in 2015 in which he showed it was possible to identify which transactions belonged to a single person but additional steps are necessary to identify that

individual (Berinato, 2015). "We didn't try to put names on it, but we know basically what you need to do that."

If the anonymity of this data isn't guaranteed then what is to be done? Schwartz and Solove (2014) describe policy regarding personally identifiable information(PII) in the United States and the EU thusly:" The U.S. approach involves multiple and inconsistent definitions of PII that are often particularly narrow. The EU approach defines PII to encompass all information identifiable to a person, a definition that can be quite broad and vague." To better address the possibility of anonymous data being reidentified they propose a change in the idea that information is either identifiable or not identifiable. They propose to add a category called "identifiable data" to which data that has "some non-remote possibility of future identification." would be categorized. They go on to cite an FTC Staff Report (FTC, 2012) to demonstrate practical policy for maintaining privacy. "First, the company must use reasonable means to ensure that the data is de-identified, or cannot be tied to a specific consumer. Second, the company must publicly promise that it will use the data only in de-identified fashion and not attempt to re-identify it. Finally, if the company makes the de-identified data available to other companies, it must contractually bind the other entities to avoid re-identifying the data and to engage in reasonable oversight to monitor compliance with these contracts"

Others believe the answer lies in individuals owning their data. Greenwood et al. (2014) present the idea that "You have the right to possess data about yourself. Regardless of what entity collects the data, the data belong to you, and you can access your data at any time. Data collectors thus play a role akin to a bank, managing data on behalf of their 'customers'." They also state that the individual should have the "right to full control over the use of your data."

Meaning a person must be able to withdraw data from a company if they want. Furthermore, they submit that an individual must have the right to dispose of or distribute their data.

Raj Sharma is another who advocates for patients having more control of their medical data. He is a co-founder and CEO of Health Wizz, a secure mobile platform for consumers to aggregate, organize and share health records based on blockchain technology. In Sharma (2018), he asks "Why shouldn't individuals own their medical records? After all, these medical records contain their personal health information and were created for them. Lab work is literally a part of the patient -- why should other people own that?" The app his company had developed is focused on individuals being able to access and control their medical data rather than wresting ownership rights away from medical institutions such as doctors' offices and hospitals where federal and state law currently places it. According to the website ,www.healthwizz.com, the app helps aggregate data from medical records from physician offices, hospitals, lab reports and data from wearables and other fitness apps. The user can then choose what information to share with whom. This app seems empowering to healthcare consumers but how much demand is there for such a thing? Companies conducting research or marketing campaigns rely on de-identified data procured from data brokers (Advisory.com, 2018). Mr. Sharma also advocates for "regulations that prevent anyone from selling or sharing our medical data without our consent…" He believes "It will force pharmaceutical companies, research organizations and hospitals to go directly to patients and request their data." (Sharma, 2017). If that were the case, surely demand for apps like Health Wizz would increase.

Years ago, the federal government passed the Health Insurance Portability and Accountability Act of 1996. Its Privacy Rule describes how different entities should handle patient information. It forms the basis for federal law on how medical data is collected and

transferred. Since it has been enacted consumers have moved more activity to the internet and the data broker industry has grown considerably and is estimated to be a $200 billion industry (Press, 2017). In 2014 the FTC conducted an investigation of the data broker industry and produced a report that included recommendations for potential legislation to Congress (FTC, 2014). At the federal level there hasn't been a comprehensive data privacy bill nor a bill targeting data brokers passed despite numerous bills being sent to Congress. Numerous attempts have been made in recent years(Bischoff, 2018). In the last Congressional session at many bills addressing data brokers were introduced, HR6675, S2577, s2342 among others(O'Donovan & McGuireWoods, 2020) and more are on the way (Fazlioglu, 2021). Perhaps the hesitancy is related to lobbying efforts made by data brokers.

Ng and Varner (2021) compared the U.S. Senate's Lobbying Disclosure Act database with companies that are registered as data brokers in Vermont and California, as required by state law. They were able to find 25 companies whose "combined spending on federal lobbying totaled $29 million in 2020. Many of the top spenders were not pure data brokers but companies that nonetheless have massive data operations. Oracle, which has spent the past decade acquiring companies that collect data, spent the most by far, with disclosure documents showing $9.57 million spent on federal lobbying." They compare these numbers with the tech industry in aggregate at $108 million and more specifically with tech titans Facebook, Apple and Google which spent $19.68 million, $18.725 million and $8.85 million respectively, in 2020. IQVIA, a company mentioned earlier that deals in de-identifed patient data spent $255,000 in 2020. I was unable to find lobbying information on other brokers who deal in de-identified data like Optum (Optum,2017) and IBM Watson Health. Both of these companies happen to be branches of larger companies. Optum is owned by United Health Group which spent $4.1 million in federal

lobbying and IBM spent $4.62 million in federal lobbying, both in 2020. The pharmaceutical manufacturing industry is using more patient data to develop drugs (Hirschler, 2018) (PEN, 2015). The industry as a whole spent $160,671,206 on federal lobbying in 2020 with major players like AstraZeneca PLC, Bayer Corp, and Pfizer Inc. spending $3.47 million, $6.51 million and $10.87 million respectively. While most of this money may not be spent on lobbying that directly impacts data broker operations it illustrates the amount of influence wielded by companies that have a stake in the acquisition of de-identified patient data.

Some state legislatures have passed privacy bills that affect data brokers. Vermont passed H.764 in 2018 which is summarized: "This act adopts consumer protection provisions relating to data brokers, including creating a new set of definitions, requiring annual registration, requiring a data security program, and requiring further study of related issues by the Attorney General." (VL , 2018) The registry is intended to make it easier for consumers and regulators to "search information concerning entities acting as third-party brokers of credit or other data, and providing information on whether brokers have opt-out or other policies consumers might want to utilize". Also, "Christopher Curtis, chief of the attorney general's public protection office, said the bill also clarifies data security requirements for commercial entities that buy or sell personal information" (Thierren, 2018).

California also passed privacy legislation in 2018. The California Consumer Privacy Act (CCPA) is more comprehensive than Vermont's H.764. It was drafted and received enough signatures to appear on the state ballot(Harwell, 2019) and passed with the requisite majority vote. The use of the ballot system in California is a process that allows citizens a way to propose laws without the support of the Governor or the Legislature (SCDOJ). At its core it grants California consumers "The right to know what personal information is collected, used, shared or

sold, both as to the categories and specific pieces of personal information; The right to delete personal information held by businesses and by extension, a business's service provider; The right to opt-out of sale of personal information.; … The right to non-discrimination in terms of price or service when a consumer exercises a privacy right under CCPA." (SCDOJ-CCPA). The cost to businesses to comply with these new regulations are estimated to total $467 million to $16.454 billion during the period 2020 to 2030 (SCDOJ-CCPA). Generally, since the CCPA was passed it has been considered the most stringent consumer data privacy law in the states but Mary Stone Ross, working with Californians for Consumer Privacy (Hill, 2018), who helped draft the act, doesn't think it will achieve its goals. "Unfortunately, the California attorney general's office predicts that even with additional resources, they will only be able to bring three enforcement actions a year, rendering the CCPA largely toothless." (Ross, 2020). She fears industry lobbying influence will further weaken the bill: " Although the legislative deal was struck in good faith, industry relentlessly lobbied for legislation that will fundamentally undermine the CCPA, while simultaneously attempting to preempt it with equally aggressive lobbying campaigns in Washington. Last session, there were over 20 bills making their way through Sacramento that would weaken the CCPA. Thankfully, due to the efforts of privacy advocates and the courage of Assemblywoman Buffy Wicks and Senator Hannah-Beth Jackson, those efforts largely failed—for now." (Ross, 2020).

Consumer data privacy continues to evolve in California. In 2020 the California Privacy Rights Act (CPRA) was approved by voters and became law. It allows for the creation of a new agency, the California Privacy Protection Agency, to handle enforcement of the act, expands on the rights granted in the CCPA and confers new rights: "Right to correct inaccurate information, right to have personal information collected subject to data minimization and purpose limitations

and the right to receive notice from businesses planning on using sensitive personal information and ask them to stop."(PRC, 2020). Companies that are subject to it are required to comply with this new act by January 1, 2023.

Conclusion

On March 2, 2021, Virginia signed the Consumer Data Protection Act (CDPA) into law. It is compared to the CCPA (Cooley, 2021). Rippy (2021) notes that the interest for privacy bills at the state level is at an all-time high and that a multitude of bills have been introduced to state legislatures. At this time just over half of the states in the union have privacy bills somewhere in the legislative process or have been signed into law. It seems likely there will be some form of consumer privacy legislation signed into law in most states or possibly at a federal level soon. Ultimately, what effect it will have is up to the participants. The general patterns explored in this paper likely overlap with other areas in technology and hopefully contribute something, if rather insignificant, to an understanding of how science, technology and society interact.

References

Advisory.com. (2018, Apr. 10) Patient data is a hot commodity. Here's how third parties (legitimately) get ahold of it. https://www.advisory.com/daily-briefing/2018/04/10/patient-data

Allen, M.(2018, Jul. 17) Health Insurers are Vacuuming Up Details About You – And It Could Raise Your Rates. NPR. https://www.npr.org/sections/health-shots/2018/07/17/629441555/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates

Atockar. (2014, Sept. 15) Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset. Neustar. https://perma.cc/E4UH-SB4D

Bischoff, P. (2018, Nov. 27) What is the Consumer Privacy Bill of Rights. Comparitech. https://www.comparitech.com/blog/vpn-privacy/consumer-privacy-bill-of-rights/

Bennet, Holly. (2017, Dec. 18) Research reveals de-identified patient data can be re-identified. The University of Melbourne. https://about.unimelb.edu.au/newsroom/news/2017/december/research-reveals-de-identified-patient-data-can-be-re-identified

Berinato, S. (2015, Feb. 9) There's No Such Thing as Anonymous Data. Harvard Business Review. https://hbr.org/2015/02/theres-no-such-thing-as-anonymous-data

CDC. National Center for Health Statistics. Health Expenditures. https://www.cdc.gov/nchs/fastats/health-expenditures.htm

Cooley. (2021, Mar 8.) Virginia Becomes Second US State to Enact Comprehensive Privacy Law. https://cdp.cooley.com/virginia-becomes-second-us-state-to-enact-comprehensive-privacy-law/

Culnane, C., Rubinstein, B., Teague, V. (2017, Dec. 8). Health Data in an Open World. https://arxiv.org/ftp/arxiv/papers/1712/1712.05627.pdf

Fazlioglu, M. (2021, Mar. 17) The first but not last comprehensive US privacy bill of 2021. International Association of Privacy Professionals. https://iapp.org/news/a/the-first-but-not-the-last-comprehensive-u-s-federal-privacy-bill-of-2021/

Federal Trade Commission (FTC). (2014) Data Brokers: A Call for Transparency and Accountability. https://www.ftc.gov/system/files/documents/reports/data- brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf

Federal Trade Commission (FTC). (2012, Mar.). Protecting Consumer Privacy in an Era of Rapid Change: Recommendations For Businesses and Policymakers https://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers

Greenwood, D., Stopczynski, A., Sweatt, B., Hardjono, T., & Pentland, A. (2014). The New Deal on Data: A Framework for Institutional Controls. In J. Lane, V. Stodden, S. Bender, & H. Nissenbaum (Eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (pp. 192-210). Cambridge: Cambridge University Press. doi:10.1017/CBO9781107590205.012

Harwell, E. (2019, Oct. 7) What Businesses need to Know About the California Consumer Privacy Act. American Bar Association. https://www.americanbar.org/groups/businesslaw/publications/blt/2019/10/ca-consumer-privacy/

Hill, K. (2018, Aug. 20) How a Woman Disappears form the History Books. Jezebel. https://jezebel.com/how-a-woman-disappears-from-the-history-books-1828393645

Hirschler, B. (2018, Mar. 1). Big Pharma, big data: why drug makers want your health records. Reuters. https://www.reuters.com/article/us-pharmaceuticals-data-idUSKCN1GD4MM

HHS. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

IQVIA. (2021, Mar. 9) IQVIA Virtual Trial Solution Used in More than 60 Trials. https://ir.iqvia.com/press-releases/press-release-details/2021/IQVIA-Virtual-Trial-Solutions-Used-in-More-than-60-Trials/default.aspx

Koontz, L. (2015). Health Information Privacy in a Changing Landscape. *Generations: Journal of the American Society on Aging, 39*(1), 97-104. https://www.jstor.org/stable/26556107

Landis-McGrath, Alice & Ballard-Stuart, Amanda. (2020, Dec. 10) The responsible, ethical use of real-world data. IBM. https://www.ibm.com/blogs/watson-health/ethical-use-real-world-data/

Loeb, Steven.(2018, Nov. 6). What big health insurers are doing with big data. Vator News. https://vator.tv/news/2018-11-06-what-big-health-insurers-are-doing-with-big-data

Montjoye, Y., Hidalgo, C.A., Verleysen, M., Blondel, V.D. (2013, Mar. 25). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports 3*, 1376. https://doi.org/10.1038/srep01376

Montjoye, Y., Radaelli, L., Singh, V.K., Pentland, A. (2015, Jan. 30) Unique in the shopping mall: On the reidentifiablity of credit card metadata. *Science vol. 347, Issue 6221*, 536-539. DOI: 10.1126/science.1256297

Narayanan A., & Shmatikov V., "Robust De-anonymization of Large Sparse Datasets," 2008 IEEE Symposium on Security and Privacy (sp 2008), Oakland, CA, USA, 2008, pp. 111-125, doi: 10.1109/SP.2008.33

Ng, A. & Varner, M. (2021, Apr. 1) The Little-Known Data Broker Industry Is Spending Big Bucks Lobbying Congress. The Markup. https://themarkup.org/privacy/2021/04/01/the-little-known-data-broker-industry-is-spending-big-bucks-lobbying-congress

O'Donovan, A & McGuireWoods LLP. (2020, Sept. 28) Federal Data Privacy Legislation: Will It Help the US Remain Competitive in the Global Marketplace? McGuireWoods. https://www.passwordprotectedlaw.com/2020/09/federal-data-privacy-legislation/

Optum. (2017) Clinformatics Data Mart. https://www.optum.com/content/dam/optum/resources/productSheets/ClinformaticsforDataMart.pdf

Press, Gil. (2017, Jan 20.) 6 Predictions for the $203 Billion Big Data Analytics Market. Forbes https://www.forbes.com/sites/gilpress/2017/01/20/6-predictions-for-the-203-billion-big-data-analytics-market/?sh=244d7d282083

Privacy Rights Clearinghouse (PRC). (2020, Dec. 10) California Privacy Rights Act: An Overview. https://privacyrights.org/resources/california-privacy-rights-act-overview

Process Excellence Network (PEN). (2015) 6 ways Pharmaceutical Companies are Using Data Analytics to Drive Innovation & Value. https://www.iqpc.com/media/1001534/35903.pdf

Rippy, Sarah. (2021, Mar. 22) US State Comprehensive Privacy Law Comparison. International Association of Privacy Professionals. https://iapp.org/resources/article/state-comparison-table/

Ross, M.S. (2020, Jan. 3) I helped draft California's new privacy law. Here's why it doesn't go far enough. Fast Company. https://www.fastcompany.com/90444501/i-helped-draft-californias-new-privacy-law-heres-why-it-doesnt-go-far-enough

Singh, G., Schulthess D., Hughes, N., Vannieuwenhuyse B., Kalra, D. (2018, Mar.) Real world big data for clinical research and drug development, Drug Discovery Today, Volume 23, Issue 3,2018, Pages 652-660, https://doi.org/10.1016/j.drudis.2017.12.002.

Schwartz, P., & Solove, D. (2014). Reconciling Personal Information in the United States and European Union. *California Law Review, 102*(4), 877-916. http://www.jstor.org/stable/23784355

Sharma, R. (2017, Oct. 1) The Privacy Myth of De-Identified Medical Data. Medium. https://medium.com/healthwizz/the-privacy-myth-of-de-identified-medical-data-10b9678e4bea

Sharma, R. (2018, Apr 23.) Who Really Owns Your Health Data? Forbes. https://www.forbes.com/sites/forbestechcouncil/2018/04/23/who-really-owns-your-health- data/?sh=ca2a9d16d62b

State of California Department of Justice(SCDOJ). Office of the Attorney General. Ballot Initiatives. https://oag.ca.gov/initiatives

State of California Department of Justice(SCDOJ-CCPA). Office of the Attorney General. California Consumer Privacy Act. https://oag.ca.gov/system/files/attachments/press_releases/CCPA%20Fact%20Sheet%20%2800 000002%29.pdf

Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Health. 671. https://dataprivacylab.org/projects/identifiability/paper1.pdf

Sweeney, L. (2005,Jun. 15). Privacy Technologies for Homeland Security. https://www.dhs.gov/xlibrary/assets/privacy/privacy_advcom_06-2005testimonysweeney.pdf

Therrien, J. (2018, May 30) Vermont first to pass data broker regulation bill. Vtdigger. https://vtdigger.org/2018/05/30/vermont-first-pass-data-broker-regulation-bill/

Thorpe, J., & Gray, E. (2015). BIG DATA AND PUBLIC HEALTH: NAVIGATING PRIVACY LAWS TO MAXIMIZE POTENTIAL. *Public Health Reports (1974-), 130*(2), 171-175. http://www.jstor.org/stable/43775470

Tsuji, Y. (2020). Medical Big Data in Japan. *Journal of Law & Cyber Warfare, 8*(1), 153-168. https://www.jstor.org/stable/26915566

Vermont Legislature (VL)(2018) Act No. 171 (H.764) Summary. legislature.vermont.gov https://legislature.vermont.gov/Documents/2018/Docs/ACTS/ACT171/ACT171%20Act%20Summary.pdf