Data Mining in Healthcare: How Far is Too Far?

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

James Perry
Spring, 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

Signature: James Perry Date: 05/13/2021

Approved: Sharon Tsai-Hsuan Ku Date 05/13/2021

Department of Engineering and Society

Introduction

This research was motivated by an interest in data sharing in healthcare for the use of AI, but I quickly found that the real issues in the industry arise from the data that doesn't come from hospitals and whose sharing isn't regulated. In this paper I will discuss the complex nature of healthcare data, the laws that govern it, and the companies that may or may not be profiting from it. There are two main topics that need to be introduced below before the body of the research: the Health Insurance Portability and Accountability Act (HIPAA), which is the set of laws that currently regulated sharing of any protected health data, and the idea of surveillance capitalism, a term that refers to the mining and use of data generated by using free services like Facebook or Google.

In 1996 Bill Clinton signed HIPAA into law. Its main purpose was to protect individuals who were without or between jobs from losing access to health insurance. There were a lot of other smaller issues covered in the act, such as streamlining healthcare records to reduce waste, emphasizing tax breaks for medical expenses, and other small improvements to the healthcare system. One of the most important aspects of the legislation was an emphasis on replacing paper records with electronic ones.

In 2003 the Privacy Rule was added to HIPAA to protect sensitive patient data by categorizing some of it as Protected Health Information (PHI). This included "Any information held by a covered entity which concerns health status, the provision of healthcare, or payment for

healthcare that can be linked to an individual". The Privacy Rule prohibited disclosure of and PHI without express permission from the patient, as well as giving the patient the right to withhold information about their healthcare from insurance companies if their treatment is privately funded. Because much of this information was now being stored electronically, in 2005 the Privacy Rule was amended to include electronic PHI records. It concluded safeguards for access, as well as standards for secure transmission of information.

The last real change to HIPAA came in the form of the Omnibus Rule in 2013, which was essentially meant to fill some of the gaps in existing rules and regulations. It specified things like specific encryption standards, as well as more strict rules on who was actually allowed access to the data in its electronic form.

In the same way that HIPAA and the healthcare industry has changed dramatically over the past 25 years, the amount of data generated by people and how it is used has changed too. To understand this change it's important to introduce a new term: surveillance capitalism. It was coined by Shoshana Zuboff, a Harvard professor and social psychologist, to describe the methodology of companies like Google and Facebook when it comes to user data. Her book, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, details how we reached our current reality of surveillance capitalism. Below is a paraphrased section of her book on the rise of Google, and it will serve as an introduction to the topic (Zuboff, 2019).

This term surveillance capitalism links the technological revolution that started in the 1990s to the industrial revolution. Both caused significant changes to the way society functioned,

and both depend on exploiting unregulated resources. The key difference is that the technological revolution was built on exploiting data generated by human behaviors. Google is the poster-company to this movement similarly to the way that Ford and General Motors were the face of the mass production movement in the early 20th century.

One of the main reasons that Google was able to distance itself from other search engines in the 90s was the way it used the data it collected from users. Every user interaction with a website generates data exhaust, a term that encompasses query spelling, grammar, search patterns, dwell time, clicks, and location. Engineers at Google were able to take that data and feed it back into their system using artificial intelligence (AI) to make a recursively updating search engine that became more accurate and intuitive over time. The problem was that after the dotcom crash of 95, investors started pushing companies to show evidence of long-term economic growth and sustainability. Up until this point Google had almost religiously avoided using paid advertisements on their site, and their ad division consisted of about 10 people. When investors started to wonder if the business plan could live up to the technology, Google made the decision to change their philosophy to meet financial expectations.

Targeted advertising was not a new concept at this point, and other search engines had been doing it for a long time. However, Google decided that they would not let the advertisers be involved in the process at all: instead, Google would choose the key words for their product. Originally, this was a simplification of the process spurred by a dislike of targeted advertising. Soon, however, because of their massive amount of user behavioral data Google was able to

move away from using keywords at all. Instead, specific ads would be targeted to individuals based off of the behavioral data that they generated.

This change in online advertisement was the tipping point that kicked off surveillance capitalism. Google broke into new territory by mining user data to increase their profits: they found new resources to exploit for gain, and ever since companies like Facebook, Amazon, and Apple have followed suit.

Literature review

One of the main goals of HIPAA from the start was to protect PHI for patients by establishing security measures. These measures govern the access and transfer of those records, but these strict rules can actually result in more problems for the hospitals. A study by Shrivastava et al. (2021) looked into how security regulations "impacted patient health information interoperability at technical (TI), semantic (SI), and organizational (OI) levels within the hospitals." The study found that these measures were often linked to higher SI and OI problems if instituted on a regional and institutional level, while hospital-wide systems were much less likely to experience TI and SI problems. While the study fails to actually address these issues in the context of big data and companies like Google, the issue it outlines should still be considered in the conversation. PHI documents still make up a large percentage of healthcare data, and the way that they are shared is vital to the system.

Insurance companies are one of the biggest stakeholders in the healthcare industry, and access to data like Google and Facebook possess would give them a huge advantage in knowing

how to evaluate risk. However, this involves using someone's personal information to potentially gain a financial advantage over them. A study done on private Australian insurance companies looked into the ethical side of the problem (Al-Saggaf, 2015). It found that this use of personal information is unethical for multiple reasons: First, there are gaps in Australia's laws regarding personal data use that leave customers open to being at a disadvantage without their consent. Second, there is potential for customers to be classified into certain groups like one that would default on payments or are likely to develop a certain disease. While this study does touch on important points like user consent, it does not deal with the real issue of data collection being done on the scale of companies like Google and its implications.

Companies outside the healthcare industry like Google or Amazon still have a lot to offer in the research field into new healthcare technologies. However, this would potentially mean giving the companies access to very personal health records for research purposes. An article on this situation in the UK looks at how to build relationships between nation health organizations and private companies to maintain trust and privacy while still improving healthcare (Horn & Kerasidou, 2020). The conclusion is to emphasize limited access, as well as transparency into how an individual's data is being used. However, while these are important points and should also be emphasized in the current situation in the US, it's also important to note that the healthcare system in the US is also privately owned and profit driven to some extent, and that we have a current set of rules under HIPAA that need to be followed.

While there has been some research done into both the interactions of data companies and the healthcare system, a key issue left out is the intersection of these companies and the massive

amount of unprotected data being generated outside of the healthcare system. This data, constantly being generated and collected through user interaction, has huge potential to affect individuals both negatively and positively.

STS Framework

The research discussed in this paper will be looked at through the framework of surveillance capitalism. The idea of surveillance capitalism is a relatively new framework through which to view data rights and privacy, but problems arise when it is applied to healthcare data and HIPAA specifically. HIPAA was written into law in the late 90s as a way to protect patients and their sensitive data, but since then, it has been changed and updated over the years to keep up new technologies and data practices. However roughly the same period of time companies like Google, Facebook, Amazon, and Apple ushered in the time of surveillance capitalism.

To get a better understanding of how surveillance capitalism is being used in the grey area of personal data privacy, it's important to look at a practical example: In April of 2018 Facebook CEO and creator Mark Zuckerberg came before Congress and the Senate to defend his companies use of user data. One of the key points discussed was the use of something called Shadow Profiles. These profiles are how Facebook collects data about individuals who do not actually use their website so that if you ever do, they already have a more complete idea of who you are. One of the ways that this data is used is the People You May Know feature for new users. Gizmodo's Kashmir Hill studied these shadow profiles and compiled stories of users who felt their privacy was violated by the algorithm (Hill, 2017). For example, an attorney she

interviewed said that he deleted Facebook after People You May Know recommended someone who he had only ever communicated with through his work email, leading him to think that Facebook was somehow monitoring an account he had never connected to it. In reality, he was more likely listed with that email address by someone who shared their contact information with Facebook. However, this example still highlights the extent of knowledge that companies like Facebook have compiled about users without explicit consent.

Data Analysis

First it is important to show that there is in fact enough data currently being collected outside of HIPAA protection that it warrants consideration for ethical issues. This data can be broken up into two categories: health information from wearable devices, and behavioral data gathered from web usage called Social Determinants of Health (SDoH). The amount of data in these two categories is massive, and probably contains much more information than is generated through hospitals. An estimate by McKinsey claims that "the average patient will, in his or her lifetime, generate about 2,750 times more data related to social and environmental influences than to clinical factors" (Singhal & Carlton, 2020).

Wearable devices have become more and more popular over the last few years and the market will continue to grow. Market and Market estimate that the market for Mobile Health Solutions will increase from $50.8 Billion in 2020 to $213.6 in 2025 (Market and Markets, 2020).

There are of course numerous benefits to these devices like increased awareness of health and fitness measurements. Some devices like the Apple Watch were also shown to be able to predict COVID-19 infections before nasal swabs through monitoring heart rate vulnerability. The question is how all of the data mentioned above interacts with the current rules set out by HIPAA. A study by Banerjee et al. (2017) lays out the different scenarios that wearable device information can be collected, and how each relates to HIPAA Privacy Rules: Information being shared between Covered Entities (CE) is protected under HIPAA and any violation can incur legal recourse. This includes both PHI and DHI records. Covered entities come into play with wearable device data when they contract third-parties like Apple to collect data for them. This means that third-party companies can store both PHI and DHI, but they violate HIPAA if they share PHI. However, data sharing between Non-Covered Entities (NCE) is not covered under HIPAA and so there is no legal recourse for potential harm. This scenario has the most potential danger because it involves uncontrolled sharing of PHI. To summarize: data collected by an NCE, unless collected through a CE contracting a third party, is not covered by HIPAA and can be shared freely, even while identifiable.

Wearable device users can see their data being recorded as they use it, and while they might not know if or how the parent company might use this data, its presence is obvious. However, this is not the case for SDoH data. This data includes details like living, working, and aging locations, level of education, socioeconomic status, and race. It can be generated from a variety of sources: web searches, location, credit card usage, and other interaction with

companies like Google, Facebook, and Microsoft. This means that an incredibly large amount of data with health implications is in the hands of Silicon Valley.

There are definitely some very positive outcomes to this data collection. For example, one study Papadopoulos et al. (2020) was able to use passively collected data from smartphone touch screens to accurately identify Parkinson's patients. This was done by using two data sources: tremor when touching the screen, and duration of touch on keys while typing. The data was used to train a deep-neural network machine learning algorithm, and it's potential to change the way Parkinson's is huge.

However, because HIPAA does not have any control over how private companies use the data they collect, they are free to research and experiment with how that data can be used for predictions. Facebook has used SDoH for a variety of applications, and not all of them have been well received. For instance, when Facebook first implemented its live stream function, many troubled individuals started to post suicide attempts. To counteract this trend, Facebook started monitoring these posts and flagging potentially problematic ones. Flagged instances were then sent to a team to review and notify police. However, because of the lack of oversight into operations like this, it is left entirely up to Facebook to train the team and set standards for flagging posts. The general reception of this plan was negative, as many thought that Facebook should not have the power to send police into people's homes as this could potentially violated the 4th amendment which protects against warrantless searches. Facebook was also in hot water when leaked documents showed that they shared personal data about users to advertisers that included mental health indicators, giving advertisers psychological insights for targeted ads

(Levin, 2020). Facebook responded in a number of ways to this accusation before ultimately claiming that the research was done to see how people express themselves, and that it was taken out of context.

Another area of concern is the de-identification of PHI records. Under the HIPAA Privacy Rule, there are two methods by which PHI records can be de-identified (Rights, 2015): First, an expert in the field can review the information and certify it as de-identified. Second, all data that can be directly associated with the individual have been removed. This means removing name, data of birth, address, relatives, contact information, etc. form the record. If these records are then re-identified either by a CE or a private company in partnership with one, they immediately return to PHI status and are protected by HIPAA. The problem arises when these de-identified records are in the possession of a company like Google. In 2019 a former University of Chicago Medical Center patient sued both the University and Google for what he believed to be a violation of HIPAA (Dyrda, 2020). The reason for this potential violation was the belief that a company like Google could easily reverse the de-identification process, giving them access to personal health data. The lawsuit was eventually dismissed because of a lack of evidence, but the issue it brings up is still valid. With the amount of data that Google and others possess, no records are ever really de-identified, and while reversing this process without consent violates HIPAA, it is important to note that is it possible, and the companies with the data to do it are partnering with hospitals more and more.

Discussion

The data presented above shows evidence of the staggering amount of data being collected and processed, as well as some of the harmful and potentially unethical ways that this data is being used. However, none of these companies have yet been shown to have broken any laws. Therefore, the question must now be asked: is this process of surveillance capitalism unethical, and if so, how can it be corrected by a system better than HIPAA is currently doing?

The most obvious ethical concern is financially exploiting people by using their own personal information without consent. However, can information taken from interacting with a free website or service technically be classified as personal? Technically the answer is no: when a user agrees to use a product like Google or Facebook they are also agreeing that the company can track how it is used. At first this made sense, as Google was using that type of information to improve its user experience. However, that original definition has changed in two fundamental ways that now make it ethically challenging: First, no matter how the data is being collected, it is hard to argue that it is no longer personal data. It is clear from the evidence that these companies have significant amounts of information on most individuals, much of it with very personal implications like physical and mental health. Second, this data is no longer being used as an exclusively mutually beneficial resource, but instead is used to target ads and potentially alter behavior. Therefore, this data can no longer be considered solely as a resource for companies with the access to exploit it, but it should also be classified as personal data and the individual should have some control over its use as such.

The California Consumer Privacy Act came into effect in 2019 and was the first law to really address the problem of controlling user data. Its purpose was to give users more control

over their data, including the ability to view it or request that it be deleted. However, after the law came into effect the initial results were inconsistent at best. According to the Wallstreet Journal, “Some companies have incorrect information on their websites about how the law affects them and consumers. Most companies acknowledge requests with emails or text messages, while other requests seem to disappear once filed. And once obtained, the volumes of data create a new burden for consumers — how to manage it” (Bensinger, 2020). This law shows progress in the fight for more control over personal data, but clearly there is still a lot of work to be done. Other states have since attempted to implement similar laws to varying degrees of success and a comprehensive breakdown can be found by International Lawyers Network (2020).

Conclusion

It seems clear that HIPAA is no longer sufficient to be the only guiding force for medical data. Data with medical implications that is gathered through user interaction cannot be separated from other data. At this point all behavioral data is health data and for its use to be regulated, all such data needs to be regulated. While the current laws in California and other states are a step in the right direction, the real problem lies in how society views these companies, and how they view us. While we may see companies like Google or Facebook as useful tools and free services, they see society as a massive resource of data that needs to be mined and exploited and we let them do it. There are a few reasons that I would say this happens, and some of these are from my own experience with these platforms: First, people do not know how much data is being mined.

Like mentioned above, downloading personal data from a company can be a huge task, and most people are not equipped to monitor it all. Second, the implications and potential harm of this data is not widely known. As shown in my research above, we are just now seeing the extent to which this data can be used to model and predict human behavior, and ever what is found is not widely spread knowledge. Lastly, there are not many alternatives. The majority of the research done for this paper was found through searching Google because it is the best platform. Similarly, the reason that Facebook and Amazon have so much traffic is because they are the best at what they do.

All the above reasons are the product of a lack of information. Society needs to realize what is happening and take control of it. A good way to look at it would be the exploitation of natural resources like oil and fossil fuels. We realized the potential danger of unregulated usage and the harm it was causing, so laws and regulations were set in place and society as a whole started to move away from it towards more sustainable methods. While this may or may not have been effective enough (that question is outside the scope of this paper), the methodology behind it is the point: laws like The California Consumer Privacy Act are just the start; people need to take control of their data and force the companies that generate it to be accountable for how they use it.

References

Al-Saggaf, Y. (2015). The Use of Data Mining by Private Health Insurance Companies and Customers' Privacy. *Cambridge Quarterly of Healthcare Ethics*, *24*(3), 281–292. https://doi.org/10.1017/s0963180114000607

Banerjee, S. S., Hemphill, T., & Longstreet, P. (2017). Wearable devices and healthcare: Data sharing and privacy. *The Information Society*, *34*(1), 49–57. https://doi.org/10.1080/01972243.2017.1391912

Bensinger, G. (2020, January 21). *So far, under California's new privacy law, firms are disclosing too little data — or far too much*. Washington Post. https://www.washingtonpost.com/technology/2020/01/21/ccpa-transparency/

Dyrda, L. (2020, September 8). *Data-sharing lawsuit against U of Chicago Medical Center, Google dismissed*. Beckers Hospital Review. https://www.beckershospitalreview.com/healthcare-information-technology/data-sharing-lawsuit-against-u-of-chicago-medical-center-google-dismissed.html

Hill, K. (2017, November 7). *How Facebook Figures Out Everyone You've Ever Met*. Gizmodo. https://gizmodo.com/how-facebook-figures-out-everyone-youve-ever-met-1819822691

Horn, R., & Kerasidou, A. (2020). Sharing whilst caring: solidarity and public trust in a data-driven healthcare system. *BMC Medical Ethics*, *21*(1). https://doi.org/10.1186/s12910-020-00553-8

International Lawyers Network. (2020, February 13). *States Are Proposing Their Own CCPA-Like Privacy Laws*. JD Supra. https://www.jdsupra.com/legalnews/states-are-proposing-their-own-ccpa-55449/

Levin, S. (2020, July 1). *Facebook told advertisers it can identify teens feeling "insecure" and "worthless."* The Guardian. https://www.theguardian.com/technology/2017/may/01/facebook-advertising-data-insecure-teens

Market and Markets. (2020). *mHealth Solutions Market*. Forecast to 2025 | By Apps, Connected Devices & Services | MarketsandMarkets. https://www.marketsandmarkets.com/Market-Reports/mhealth-apps-and-solutions-market-1232.html

Papadopoulos, A., Iakovakis, D., Klingelhoefer, L., Bostantjopoulou, S., Chaudhuri, K. R., Kyritsis, K., Hadjidimitriou, S., Charisis, V., Hadjileontiadis, L. J., & Delopoulos, A. (2020). Unobtrusive detection of Parkinson's disease from multi-modal and in-the-wild sensor data using deep learning techniques. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-78418-8

Rights, O. F. C. (2015, November 6). *Methods for De-identification of PHI*. HHS.Gov. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard

Shrivastava, U., Song, J., Han, B. T., & Dietzman, D. (2021). Do data security measures, privacy regulations, and communication standards impact the interoperability of patient health information? A cross-country investigation. *International Journal of Medical Informatics*, *148*, 104401. https://doi.org/10.1016/j.ijmedinf.2021.104401

Singhal, S., & Carlton, S. (2020, March 1). *The era of exponential improvement in healthcare?* McKinsey & Company. https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-era-of-exponential-improvement-in-healthcare

Zuboff, S. (2019, October 1). *How Google Discovered the Value of Surveillance*. Longreads. https://longreads.com/2019/09/05/how-google-discovered-the-value-of-surveillance/