Patient Data Privacy: HIPAA, the Failure of Anonymization, and Suggested Solutions

Mary Ziemba

I. Introduction

SAY, for example, John is a young man who used to be an alcoholic. He successfully completed treatment at a drug rehabilitation center and turned his life around. After treatment, John applies and interviews for a job, and his potential employer is impressed with his skills and professionalism in the interview process. For obvious reasons, John would not want his potential future employer to know about his time in rehab. If that fact were made known to the employer, Johns chances at getting the job might be jeopardized. Situations like Johnsamong many othersshow how important it is to protect a patients private medical records, and how serious the repercussions of exposing such records could be.

Properly protecting patient privacy is a problem that is not only social and ethical, but also inherently technical, given the security and privacy know-how needed to protect patient data. The widespread use of EHRselectronic health recordsin medicine demands that encryption, user authentication, and other highly-technical skills are used to protect patient privacy and security. Despite this necessity, a host of social and political factors have caused serious breaches of patient privacy and security. HIPAA, the United States primary means of protecting patient privacy, maintains an inadequate definition of patient privacy that leaves individuals in public health datasets vulnerable to de-anonymization attacks. The high cost and complexity of robust systems, the decentralization of EHRs, and the human element of healthcare present real threats to patient data privacy and security.

To the average American, it may be surprisingly easy to re-identify data that is ostensibly anonymous. Re-identification is the practice of matching de-identified data with publicly available information, or auxiliary data, in order to discover the individual to which the data belongs to [1]. Here, I explore several technical and ethical concerns related to a serious threat to the privacy of EHRsre-identification and de-anoymization. I explore the technical issues in Title II of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), known more commonly as the HIPAA Privacy Rule. I argue in support of several changes that will protect the privacy of patients in a more robust way.

II. HIPAA and Patient Privacy

The Health Insurance Portability and Accountability Act, better known as HIPAA, was passed under President Bill Clinton in 1996. It was a landmark law in the medical field, introducing sweeping changes to the practice of recording and disclosing patient data. Among the many changes it made to healthcare administration, including changes in health insurance portability and group health plans, it is most noted for its new requirements surrounding the privacy and disclosure of patient data and medical history.

In 1996, when the law was passed, the Internet was still nascentits users and lawmakers still defining the ways it could be used and the proper ways for it to be used. When HIPAA was written, Hotmail was just released, and AOL Instant Messenger was just reaching popularity. Virtually no doctors offices and hospitals used electronic health records (EHRs) to track patient care. Needless to say, when the Health Insurance Portability and Accountability Act, better known as HIPAA, was passed under President Bill Clinton that year, it was not prepared to withstand the changes to medicine that computing and the Internet would bring.

The rise of the Internet and electronic patient data collection presents two major challenges. First is the portability of EHRs between different IT systems. Many hospitals and doctors offices rely on EHR systems such as Cerner, Meditech, and InterSystems instead of filing cabinets and folders to store patient information. The systems themselves were designed for this purposenot to make it easy for different providers to transfer EHRs to one another. But when a patient moves from one provider to another and a must view the patients medical history, getting the patients EHR from a hospital with a different IT system can be difficult to impossible.

The second major problem is the protection of patient privacy. Title II of HIPAA, better known as the HIPAA Privacy Rule, places strict regulations on the use and disclosure of patient data [2]. It is meant to give patients control over the disclosure of their health information, and sets guidelines for data usage when a patient is medically unable to consent to do so. It is also meant limit the use of protected health information, or PHI, that might compromise a patients private medical history if the PHI were released. Additionally, it defines circumstances under which HIPAAs restrictions on the release of PHI may be used without patient consentin the interest of public safety or criminal justice, for example.

This is a final paper for COMPSCI 342S: Technical and Social Analysis of Information, Duke University, Professor Salman Azhar, Fall 2017

The protection of PHI is an extremely important responsibility of HIPAA; however, the Privacy Rule was, as mentioned, written at a time before computing and the Internet were as pervasive as they are today. Much of the content of the HIPAA Privacy Ruleand its subsequent revisions in 2000 and 2002 [3]are concerned with proper procedure after a breach of health data, or permissions needed to release information to entities such as health insurance companies and family members. But as computer scientists learn more and more about data privacy, however, other threats to the privacy of patients PHI have been revealed. In the below sections, I discuss the threat of data re-identification to patient privacy, and how HIPAA and other privacy laws might be improved to better protect against re-identification.

III. Data re-identification

Here, I will discuss the process of data re-identification, which will bring to light several issues related to the privacy of health data.

As stated earlier, data re-identification is the practice of matching de-identified data with publicly available information, or auxiliary data, in order to discover the individual to which the data belongs to [1]. Data re-identification shows the failures of anonymizing data, and some other methods by which data owners inadvertently release information about individuals in a dataset.

A simple example best explains data re-identification. Say that the rehab center maintains a database of the EHRs of all its patients, past and present. A simplified view of the database might look like this:

Race	Birth date	sex	zipcode	treatment
Asian	6/11/1966	Female	13090	alcohol
White	10/18/1975	Male	29483	amphetamines
Black	6/26/1962	Male	19125	alcohol
Asian	11/10/1989	Male	60067	alcohol
Hispanic/Latino	3/23/1966	Male	90210	cocaine
Hispanic/Latino	9/23/1965	Female	65715	prescription drugs
Hispanic/Latino	12/21/1983	Female	11510	prescription drugs
White	2/2/1988	Female	96815	cocaine
White	3/6/1976	Female	60185	amphetamines
American Indian	9/5/1968	Female	56001	alcohol

Now say that another healthcare provider releases this dataset:

name	birthdate	sex	zipcode	smoker?
Taylor	11/10/1989	Male	60067	yes
Ashley	3/6/1976	Female	60185	yes
Kevin	6/26/1962	Male	19125	no
Elizabeth	6/11/1966	Female	13090	no

If one were able to obtain both of these datasets, she could form the following table:

Name	Race	birthdate	sex	zipcode	treatment	smoker?
Asian	Taylor	11/10/1989	Male	60067	alcohol	yes
White	Ashley	3/6/1976	Female	60185	amphetamines	yes
Black	Kevin	6/26/1962	Male	19125	alcohol	no
Asian	Elizabeth	6/11/1966	Female	13090	alcohol	no

This procedure, known as an inner join between two tables, allows someone to construct a database that reveals more about each individual in the dataset than either intended.

An inner join such as this one relies on a surprising fact about the American populationthat the combination of an individuals birthdate, gender, and zip code is unique for about 87 percent of Americans [4]. So while it is not guaranteed that the table above is accurateit could be the case that there is, say, another male born on November 10, 1989 in the 60067 zip codethe fact has the potential to be dangerously revelatory. Whats more, individuals in sparsely-populated zip codes might be identified by even less than the combination of their zip code, gender, and birthday. For example, in one Charlotte, North Carolina zip code with a population of only ten individuals, there is only one fourteen-year-old boy. He can be uniquely identified by only zip code and a four-year range for his birthday [6].

It might seem unlikely that such revealing datasets as the ones used in this example would be publicly available in the first place. But nowadays, data scientists make a fairly pessimistic assumption about the availability of auxiliary information to identify individuals in a dataset due to the proliferation of data on the Internet. People innocently make revelatory social

media posts, creating a veritable online diary of millions of people. It may take only one leakage or inadequate anonymization of sensitive datasetsuch as the one containing information about Johns time in rehabto be destructive to a persons privacy.

IV. Two revealing examples of the failure of data anonymization

Two recent studies reveal the potentially dangerous outcomes of anonymization of data.

A. Matching Known Patients to Health Records in Washington State Data

A 2013 study by Dr. Latanya Sweeney was a instrumental in revealing the failures of data anonymization [8]. In her study, Sweeney hypothesized that publicly available hospitalization data from the state of Washington could be re-identified using newspaper articles about hospitalizations.

In the study, Sweeney obtained a dataset of nearly every hospitalization in Washington state during the year 2011. The data included the patients age in years and months, zip code, and symptoms, as well as the hospital, attending doctor, and date of the hospitalization. Sweeney also obtained 81 newspaper articles published in the state that year that used the word hospitalized.

Sweeney took information from the newspaper articlethe patients age, residence, and symptomsand attempted to identify their database records. She could definitively link 35 of the individuals in the newspaper articles with their corresponding records of the database. She confirmed her findings with the patients themselves via the journalists.

The study shows the failure of anonymization in keeping patient data private. Any adversaryperhaps a creditor seeking repayment, or a blackmailercould use Sweeneys techniques to find out private medical information about someone. In a situation like the one described in the beginning of this articlewhen a persons livelihood or reputation might be on the linethe misuse of public health data could be devastating.

B. Genomic data and the danger of trail re-identification

Trail re-identification presents another serious threat to patient privacy. In trail re-identification, an adversary independently reconstructs the trails of locations that identified entities and their un-identified data visited, which can then employed for re-identification via trail matching [9].

Trail re-identification is best explained with an example. Drs. Sweeney and Bradley Malin performed an experiment of trail re-identification of genomic data collected at various hospitals [10]. The researchers used individuals in a publicly-available genomic database collected in the state of Illinois between 1990 and 1997. Patients, who had one of several genomic disorders such as cystic fibrosis and Huntingtons Disease, had their genomic information collected at several hospitals for treatment purposes.

Patients would leave DNA samples at several hospitals, who would record the genetic information along with some identifying information about the individual. The hospitals released the data as parts of longitudinal studies, with some identifying information about the patients removed. Sweeney and Malin would search for each patients unique DNA sequence in several hospitals databases and match up ones that were determined to be from the same individual. Using the auxiliary data from each database, which might have included age or zip code, Sweeney and Malin could definitively re-identify about 58% of the individuals who had left their DNA in one of the hospitals databases.

Trail re-identification is the process of identifying an individual across datasets by collecting the auxiliary information at each source. Linking a persons name to their public genetic information is a scary proposition for many, and could lead to malicious activity by adversaries. It is especially unfortunate that individuals with genetic disorders, who are more likely to leave genetic data at a hospital, are more susceptible to trail re-identification using DNA.

V. Current attempts to prevent re-identification

Although the failures of data anonymization are numerous, the idea that anonymization is a safe way to protect an individuals data is still prolific. The Washington state data used to identify 38 individuals health records in the aforementioned study is still publicly available in the form that the researchers encountered it [11]. Inadequate knowledge of the dangers of anonymization has led to inadequate legal protections of patient health data.

The HIPAA Privacy Rule should perhaps the best line of defense against inadequate anonymization. However, certain aspects of the wording and implementation of the HIPAA Privacy Rule make health data prone to re-identification attacks.

A. Problems with the safe harbor provision

The authors of the HIPAA Privacy Rule expressed the importance of anonymization of health records in public datasets by creating rules surrounding DHIthe de-identification of health information. The authors of HIPAA left the exact definition of what constitutes DHI to the Department of Health and Human Services (HHS) [5].

Data can meet one of two criteria in order to be considered sufficiently private under HIPAA. First, there is the expert determination method:

> A covered entity may determine that health information is not individually identifiable health information only if a person with appropriate knowledge and experience determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information. [12]

Given the huge amount of health data that exists and the relatively small number of statisticians available to personally analyze it, the expert determination method is less popular.

The more popular way of de-identifying information comes via the safe harbor standard. If all of the following identifiers of are removed from a dataset record, the data complies with the HIPAA Privacy Rule:

- Name
- Address (all geographic subdivisions smaller than state, including street address, city, county, or ZIP code)
- All elements (except years) of dates related to an individual (including birth date, admission date, discharge date, date of death, and exact age if over 89)
- Telephone numbers
- FAX number
- Email address
- Social Security number
- Medical record number
- Health plan beneficiary number
- Account number
- Certificate/license number
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers or serial numbers
- Web URLs
- IP address
- Biometric identifiers, including finger or voice prints
- Full-face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code [12]

The list provided by the HIPAA privacy rule is fairly extensive, and certainly reflects a desire on the part of HHS to protect patients individual privacy. But the list also leaves little room for interpretation, and cannot be added to in the event that other factors are discovered to be identifying. Especially given the rate at which scientific research moves forward, the law is not particularly robust to the ever-growing list of ways to re-identify data. Even a database that does not include any of the 18 identifiers listed in the safe harbor standard, but is deemed non-private by a statistician, could still be released legally.

B. The failure of data usage agreements

Other medical data distributors depend on a data usage agreements to protect their datasets from re-identification. In order to obtain the Washington state hospitalization data mentioned earlier, data requestors were only required to sign a contract with the distributor, the Healthcare Cost and Utilization Project (HCUP), stating that the obtainer would not misuse or attempt to re-identify any individuals in order to obtain the data [7]. HCUP also manages the hospitalization data from many other states, protecting it with the same data usage agreement [11]. Given the high chance of re-identification in this hospitalization data, I contend that protecting such sensitive data with a mere contract is inadequate. HCUP and other similar distributors should not trust that all data requestors will not re-identify individuals just because they signed a contract that they would notthis is like putting a band-aid on a bullet hole. Privacy researchers see data usage agreements as last resorts to protect data privacyfar from the best practice of ensuring that release mechanisms of data do not reveal any sensitive data in the first place [13]. Later in this paper, after discussing the ethical concerns of balancing patient privacy and medical research, I will discuss in more detail such mechanisms.

VI. Ethical concerns

Maintaining patient privacy is difficult and important work. But privacy preservation does not come without cost. HIPAA has the very difficult job of balancing the personal privacy concerns of individuals with the importance of using data in medical and social research.

The individual right to privacy has long been an important part of American law [14]. Before the age of big data, the right to privacy mainly allowed an individual to protect his or her right to their public image from unwanted exposure or disclosure of private or embarrassing facts. However, since data collection and distribution has proliferated since the dawn of the Internet in the 1990s, the nature of the right to privacy has been the subject of much debate. Given that cell phones track an individuals location, search engines track nearly all internet traffic, and wearables track a persons heartbeat, one might wonder if Americans are slowly becoming complacent about their right to privacy. Medical data, however, is of considerable more concern than a users browsing habits. Americans have reason for concern especially given the often-weak protection of health and medical data.

Their concern, however, sits in diametric opposition to researchers desires to make meaningful analyses of medical data. Researchers must often jump through hoops that HIPAA has set up in order to even access data, much less be able to draw meaningful conclusions from the data that lead to better patient outcomes. Wanting to be able to analyze data in any way they choose, these researchers, with good reason, tend to be less interested in the personal privacy of individuals in a database.

Paul Ohm perfectly summed up the difficult tradeoff between individual privacy and data usability: Data can be either useful or perfectly anonymous but never both [5]. I contend that some middle ground must be found between anonymity and usability of data. Copious amounts of research at Americass leading institutions is being done on differential privacy of data, including several privacy-preserving methods that more reliably protect individual privacy while maintaining utility of the dataset. The REIDIT algorithms, for example, attempt to prevent trail re-identification of data [9]. And Dr. Latanya Sweeneys work in /k/-anonymity provides a way to ensure that at least k individuals in a dataset cannot be distinguished from one another [15]. Given the importance of medical data to patient care and the dangers that can result from its improper exposure, data privacy researchers should work more closely with medical practitioners to achieve a fairer balance between the usefulness of datasets and the privacy of the individuals in them.

VII. Recommendations

All things considered, I recommend a few technical and policy changes be made to data released under the HIPAA Privacy Rule:

1) Do not rely on data privacy agreements to protect privacy: First, I recommend that data distributors do not rely primarily on data privacy agreements in order to maintain patient privacy. Instead, a dataset should be rigorously reviewed by a statistician or other scientist similarly well-versed in data privacy methods. This should help minimizeand ideally preventthe release of data that is highly re-identifiable, such as the DNA or hospitalization datasets discussed above. Because robust research on more private database storage mechanisms exists and is fairly easily applicable, a dataset should first be mathematically protected against privacy attacks, then protected by a data usage agreement to help ensure that malicious action does not take place.

2) Repeal HIPAAs safe harbor provision: Closely related to my previous suggestion, repealing HIPAAs safe harbor provision would help prevent malicious use of publicly-available data. The absence of the eighteen identifiers is a good first step toward privacy, but is by no means a guarantee of privacy. Data protected under this provision can still be very revelatory of an individual. Further, the safe harbor provision is not the best legal approach to preserving privacy, considering how quickly scientific research and new privacy attack methods emerge.

3) Expand privacy research and knowledge: By making computer scientists and computer science students more knowledgable about the field of data privacy, there will be a greater number of qualified computer scientists who can ensure that health data is sufficiently private. Making classes in data privacy more available to undergraduates studying computer science can help ensure that enough computer scientists are qualified to understand the importance of data privacy not only in medical data, but in any especially sensitive data.

VIII. Conclusion

The study of data privacy is becoming more and more important every day, as more and more sensitive health and medical data is produced daily. On the upside, this explosion of data allows for the potential for medical researchers to use the data to advance patient care and potentially save lives. On the other hand, it can lead to the violation of privacy of individuals in the dataset, leading to potentially dangerous uses of a patients sensitive medical information. Greater knowledge of the importance of patient privacy, as well as the modification of HIPAA laws to better protect a patients individual privacy, are needed to ensure that data is both useful to researchers and protective of the individuals in the datasets.

REFERENCES

[1]
[2]
[3]
[4]
[5]
[6]
[7]
[8]
[9]
[10]
[11]
[12]
[13]
[14]
[15]