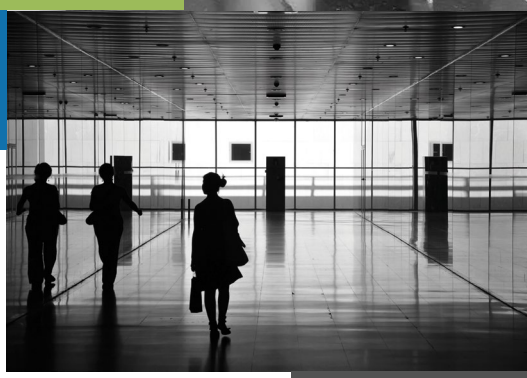
PRIVACY ANALYTICS

WHITE PAPER

HOW TO SAFEGUARD PHI

Context is king when it comes to safeguarding Protected Health Information (PHI). As patients, we share many personal details with our care providers. Concerns over who has access to this information and how it may be used can cause us as much worry as our health issues. Effectively safeguarding PHI means knowing who will have access to the data, how it will be stored and what details it contains. In other words, its context for use.



PRIVACY ANALYTICS

Nothing personal.

Context Matters

Datasets that are improperly de-identified carry a significant risk of patient privacy being compromised. The patients may be further harmed when new information is learned about them as a result of their re-identification. Safeguarding the privacy of patients is achieved through two measures: securing the data from unauthorized access and applying data de-identification.

Requiring organizations to have privacy policies and security protocols in place to protect their data reduces risk without impairing data quality. HIPAA legislation governs the disclosure of protected health information (PHI) in the U.S. and meeting the rule of law may require further steps to be taken to reduce the chances of a person being re-identified.

Context Matters: Sharing Data for Secondary Purposes

The details that can be used to describe us – our name, age, gender, and even our diseases and health conditions – are data. Together, these personal details can be used to identify who we are. When these details are gathered to inform our care and treatment as patients, the data is referred to as protected health information (PHI).

Beyond its main use for patient care, PHI can be used to conduct research, perform data analysis and undertake marketing activities, among other things. When analysts and marketers do research on health data they are not looking at the details of specific individuals. Rather, they are looking at the information of groups of individuals to identify patterns, trends and relationships that can be uncovered in the data. For this valuable research to take place healthcare providers, health insurers, governments, pharmaceutical manufacturers and other organizations that collect health information need to be able to share data for secondary purposes. This lets researchers analyze real-world data without the risk of identifying the people behind it.

Context is critical when we discuss the secondary uses of health data. Understanding the circumstances under which the data will be shared – who will have access to it and how will it be protected – is crucial to ensuring that privacy is maintained. At the same time, the data's quality must be preserved so it is useful for mining new knowledge.

This white paper examines a case where an individual was identified from PHI. We'll look at the risks to privacy when PHI is not effectively de-identified, the value in preserving data quality, and how effective the de-identification methods of Safe Harbor and Expert Determination are in safeguarding health information.

Public Data and the Risk of Demonstration Attacks

We begin this discussion by looking at the case of a Washington State man who was re-identified from his PHI[1].

In 2011, a Vietnam veteran named Ray Boylston had a motorcycle accident when he suffered a diabetic shock while riding. The incident was covered briefly in the local Washington paper (See Fig 1). The record relating to Ray's week-long stay at Lincoln Hospital was subsequently included in the hospital's inpatient database. As

Public Data and the Risk of Demonstration Attacks

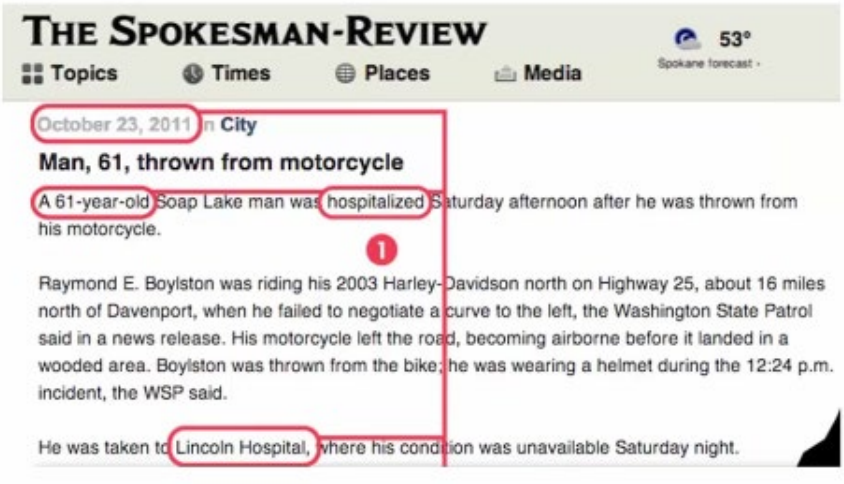
part of a larger statewide project, this database was made available for purchase. Although the information was mainly bought by researchers and insurance companies there were no restrictions on who could purchase it. In this regard, the data was publicly available.

Demonstration attacks are a type of re-identification attack done on publicly available data and that look for records that are easy to re-identify. In Ray Boylston's case, a well-meaning researcher who had access to the data wanted to show that it had not been properly de-identified before it was made available. By scanning the local news items from the area, he was able to find the report of Ray's accident in the Spokesman-Review. The report contained enough identifying information about Ray – such as his gender, age, admission date and cause of trauma – that the researcher was able to pinpoint Ray's record in the inpatient database[2].

The important thing is not that we learn Ray Boylston had a motorcycle accident; this was public information that was already available in the newspaper. However, by identifying Ray we have the potential to learn more about him from this database – his zip code, occupation and other diseases he may have. This risk is not about the information we use to identify someone; it is about the fact that we can learn more information about this person once they are identified.

Understanding the context in which this data was shared is critical to understanding this issue. This was a case of a public data release. Public data releases inherently pose a very high risk to privacy since anyone can access the data and potentially launch a re-identification attack.

Context means that we try to protect the data from re-identification attacks through other means. If security and privacy policies are put in place to control who can access the data, then we're no longer dealing with a public data release. By limiting access to the data, we put protection in place. A second way we can add protection for privacy is by de-identifying some of the data to


THE SPOKESMAN-REVIEW

53°
Spokane forecast ›

Topics Times Places Media

October 23, 2011 in City

Man, 61, thrown from motorcycle

A 61-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle.

1

Raymond E. Boylston was riding his 2003 Harley-Davidson north on Highway 25, about 16 miles north of Davenport, when he failed to negotiate a curve to the left, the Washington State Patrol said in a news release. His motorcycle left the road, becoming airborne before it landed in a wooded area. Boylston was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident, the WSP said.

He was taken to Lincoln Hospital, where his condition was unavailable Saturday night.

Figure 1: News report of Ray Boylston's motorcycle accident with identifiers highlighted

Safely Unlocking the Value of Data

make sure the risks of an individual being re-identified are very small. If data is not properly de-identified before it is shared it may still be considered to be PHI. This has significant legal implications, as many countries have legislation to regulate the use and disclosure of PHI, including the United States, Canada and the European Union. If the data is determined to be PHI and a re-identification attack on that data is successful, the organization is now dealing with a case of data breach. In the U.S., the average cost for a lost or stolen record is in the range of $208[3] to $217[4] per affected record. These data breaches also come with regulatory requirements that individuals affected by the breach be notified.

When the state inpatient database for Washington was examined, it was found that 84% of the records were at risk of being re-identified based on simple demographics such as gender, age and zip code[5]. By measuring the risk of re-identification when using health data for secondary purposes, privacy should be paramount.The next section will look at how this can be done.

> How do we maintain the quality of the data while reducing the risk to privacy?

Safely Unlocking the Value of Data

How do we maintain the quality of the data while reducing the risk to privacy? The main way to do this is by ensuring through the context that the risk to privacy is already low. This means looking at where the data will be stored and who will have access to it, then putting contracts in place that lay out the limitations on the use and disclosure of the data.

However, if the information that we are dealing with is highly sensitive and its disclosure could significantly harm the individual, we may need to go further and apply data de-identification. One way to remove PHI from a dataset is through masking. Masking data removes the obvious identifiers such as name, address and medical record number. These types of identifiers are also referred to as direct identifiers. By removing the obvious fields, masking reduces the risk of re-identification but it also acts as a blunt instrument that negatively impacts the data's quality for analysis.

Certain de-identification methods, on the other hand, can retain the high analytic quality of the data. These techniques aim to change the data as little as possible so that the data retains its granularity while still meeting the privacy objectives. These methods largely focus on the quasi-identifiers in the data; information such as date of birth, income, marital status or zip code. Quasi-identifiers, while not directly identifying on their own, can be used in combination to identify an individual. For example, if we can see that there is only one person in a given zip code who is 89 years old, it is easy to determine who that person is.

HIPAA Privacy: Two Methods to De-Identify Protected Health Information

Safely unlocking the value of data for secondary purposes means that we reduce the risk of re-identification as much as possible through the data's context so that changes to the data from de-identification methods are minimized, which in turn maximizes the data's quality.

HIPAA Privacy: Two Methods to De-Identify Protected Health Information

In the U.S., the Health Information Portability and Accountability Act's (HIPAA) Privacy Rule sets the standards for the use and disclosure of PHI. In addition, it provides a framework for the de-identification of PHI and outlines two methods to accomplish this: Safe Harbor and Expert Determination.

Safe Harbor has two requirements:

i. The removal of 18 elements from the data. Of these, 16 are direct identifiers. The other two elements are zip code and any dates[6].

ii. The second requirement is that there is no actual knowledge that could be used, alone or in combination, to identify an individual. This implies that there can be no direct knowledge that you could re-identify someone from the information left in the data.

To meet this second requirement, it's necessary to "mine" the data looking for opportunities where actual knowledge may exist. For example, if there is a variable in the data that captures occupation and you have a person with an uncommon occupation, like Senator, that is direct knowledge that could be used to re-identify an individual.

The other de-identification method under the HIPAA Privacy Rule is Expert Determination. It is given this name because it requires the involvement of a de-identification expert to study the data for relationships that exist among the variables and who can run statistical approaches to measure the risk of re-identification. This risk-based methodology requires:

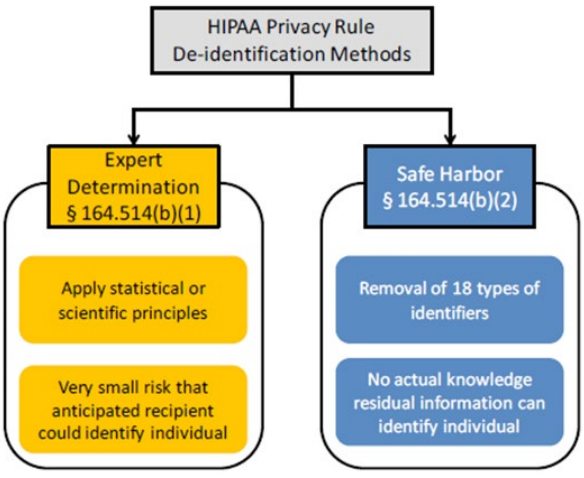


Figure 2: De-Identification Methods Covered Under the HIPAA Privacy Rule

De-Identification in Practice

i. That the risk is very small that the information could be used alone, or in combination with other available information, by an anticipated recipient to identify an individual who is a subject of the information; and

ii. Documentation of the methods and results of the analysis to justify this determination.

The variables we are looking at with Expert Determination are quasi-identifiers like gender, age and income. In order to use the data for secondary research, we want to keep the quasi-identifiers as unchanged as possible. These pieces of information can be very identifying, however, particularly when combined. It is important that we look carefully at the quasi-identifiers and the relationships among these variables to determine what is going on in the data.

The next section of this paper will look at the application of each of these methods with respect to the Ray Boylston case.

De-Identification in Practice

The first step in de-identifying data is to class the identifiers as either direct identifiers or quasi-identifiers. Direct identifiers are then masked; they have no value for the research to be conducted so are removed. As we indicated earlier, masking on its own does not do enough to guarantee the anonymity of the individuals in the data. The inpatient database released by the State of Washington was masked so that all direct identifiers had been removed, yet Ray Boylston was still found. We need to look at techniques that can be used in addition to masking.

If we look at applying the Safe Harbor method to the Washington State inpatient database, the risk of re-identification decreases from 84% of records to 33% of records . The risk is significantly reduced using Safe Harbor, but much of the analytic utility of the data is also reduced. When looking at the usefulness of data, a number of respected organizations provide frameworks or guidelines for the disclosure of de-identified health information that is both risk-based and compliance-focused. These include the Health Information Trust Alliance (HITRUST), the Institute of Medicine, PHUSE (Pharmaceutical Users Software Exchange) and the Council of Canadian Academies. As we will see in the next section, the use of a risk-based approach like Expert Determination, along with strong contextual protections, allows for data that is valuable for analysis.

Applying Expert Determination in the Case of Ray Boylston

One question with Expert Determination is what is meant by the phrase “a very small risk.” Determining what is very small is done by considering the context of the data release and looking at past precedents for guidance to define what is an acceptable level of risk. So we look at where the risk threshold was set in given cases by reputable organizations, such as the Census Bureau, when releasing data.

Unfortunately, it is impossible to guarantee zero risk; there is always a small residual risk in releasing data just as we face small risks in the activities we do every day. Even the simple act of crossing the street carries some small risk! So Expert Determination is really about risk management, where HIPAA is seeking proof that

Conclusion

CONTACT US

251 Laurier Ave W
Suite 200
Ottawa, Ontario, Canada
K1P 5J6

www.privacy-analytics.com

sales@privacy-analytics.com

shows a suitably low risk for re-identification was achieved. Once the risk threshold for the data is set, the next step is to measure the risk in the data by looking at the quasi-identifiers and at plausible attacks. When the risk in the data has been measured, we can determine if it needs to be changed to make people less identifiable. If the risk in the data is found to be above the threshold then the data will need to be transformed using de-identification techniques, such as aggregation or suppression, to bring the risk down[8].

By ensuring that basic security controls are in place, we can protect the data and then use Expert Determination to lower the re-identification risk even further. For the inpatient database for the State of Washington, it was possible to bring the risk level very close to 0%. If the data is protected with security practices and protocols then the context provides an already reduced risk so that, with limited changes to the data, it's possible to ensure a very small risk of re-identification and keep good quality data for analytic purposes.

Conclusion

When healthcare data is shared for secondary purposes, effectively protecting the privacy of patients comes down to assessing the context in which the data will be shared. First and foremost, policies and procedures are needed to lay out how the data will be protected from unauthorized use.

When we hear about cases like Ray Boylston and the State of Washington's inpatient database it may sound like sharing data for additional research puts patient privacy at risk. However, when we look at re-identification attacks on actual de-identified health data the risks are, in fact, very small. The Washington inpatient database was a public data release which inherently made the risk very high. Furthermore, the data in this situation wasn't de-identified, it was superficially masked.

Maintaining the quality of the data is important if we wish to obtain valuable insights from the analysis. Masking tools have a negative impact when it comes to sharing data for research and analysis since their approach is too broad and generic. While some implementations of Safe Harbor can preserve privacy and allow for some amount of

Conclusion

CONTACT US

251 Laurier Ave W
Suite 200
Ottawa, Ontario, Canada
K1P 5J6

www.privacy-analytics.com

sales@privacy-analytics.com

analytics, a comprehensive analysis of a wide-ranging set of variables demands the rigor of a risk-based approach like Expert Determination. When a de-identification process based on the Expert Determination method is used, we find that the risk of re-identification can be effectively eliminated and still allowing for meaningful analysis of real-world data.

To learn more about this topic, watch the webinar How to Safeguard PHI available on the Privacy Analytics website.

Sources

1. Dawes, Terry (2015, March 19). Ottawa's Privacy Analytics Participates in HITRUST Health De-Identification Framework. Cantech Letter. Retrieved from http://www.cantechletter.com/2015/03/ottawas-privacy-analytics-participates-in-hi-trust-health-de-identification-framework/

2. In this case, the researcher at the Spokesman-Review contacted Boylston and sought his consent to use these findings before publishing the results of the work.

3. El Emam, Khaled. (2013). *A Guide to the De-Identification of Personal Health Information.* Boca Raton, FL: CRC Press/Auerbach.

4. Ponemon Institute LLC. (2015, May). 2015 Cost of Data Breach Study: United States. Ponemon Institute. Retrieved from http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&htmlfid=SEW03055USEN&attachment=SEW03055USEN.PDF

5. Privacy Analytics reviewed the data and ran various risk measurements on the quasi-public dataset. The State Inpatient Database was obtained with permission from The Healthcare Cost and Utilization Project (HCUP). To learn more, visit their website: https://www.hcup-us.ahrq.gov/.

6. For more information on the Safe Harbor elements, see De-Identification 201: Fundamentals of Data De-Identification. http://www.privacy-analytics.com/de-id-university/white-papers/de-identification-201/

7. This figure is based on Privacy Analytics risk assessment testing of the quasi-public dataset using Safe Harbor. The State Inpatient Database was obtained with permission from The Healthcare Cost and Utilization Project (HCUP).

8. For more information on the techniques used in de-identification, see De-Identification 201: Fundamentals of Data De-Identification. http://www.privacy-analytics.com/de-id-university/white-papers/de-identification-201/