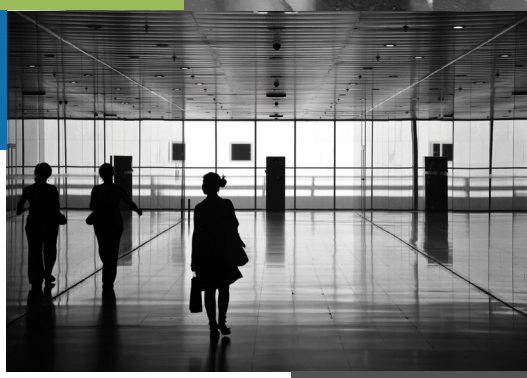
PRIVACY ANALYTICS

WHITE PAPER

De-identification 101

The growth in the use of electronic medical records, electronic insurance claims processing and other hospital software systems has led to a rise in the collection and storage of personal health information. Beyond the provision of patient care, this information can be invaluable in driving innovative research and providing new insights to address challenging healthcare problems. De-identification allows for the sharing of personal health information by protecting individual privacy. By de-identifying a dataset, the chance that a person can be identified from their data is drastically reduced.



PRIVACY ANALYTICS

a QuintilesIMS company

Protecting Personal Health Information: A Primer

Executive Summary

Some of the best standards for the privacy of personal information exist in the healthcare sector, with many jurisdictions – including the United States, Canada and the European Union – having legislation in place. Organizations need to focus on protecting their data holdings now more than ever. Risks to data privacy can come not only from external threats, like hackers and criminals, but also from inadvertent data leaks and security vulnerabilities. In addition to the legal ramifications, data beaches can also impact an organization's finances and reputation. When implementing a de-identification solution, organizations can opt for a homegrown approach, engage a de-identification expert to act as a consultant, or they can purchase commercially available software that can automate the process.

The Rise of Protected Health Information

Every day we undertake activities that generate personal information about us. More often than not, this information is now captured and stored in electronic databases. When we go to the store to make a purchase with our credit card, go online to do a search, or purchase an airline ticket to take a trip, a record is created that contains our personal details. With the growing use of electronic medical records (EMRs) and the prevalence of electronic insurance claims processing and other hospital software systems, increasingly it is our health information that's being captured. Protected health information (PHI) includes health plan numbers, hospital admissions data, disease diagnoses, prescription information, treatment plans, lab results and other data about our health and well-being that can be used to uniquely identify us.

While the primary purpose of capturing PHI is to provide patient care, it is recognized that valuable insights can come from sharing this information with researchers and analysts. Using data in this way often means making it available to outside organizations or groups.

When sharing information sources, healthcare organizations must consider how they will protect the privacy of the individuals whose information is contained within the data. The disclosure of an individual's private health information – whether intentional or not – can lead to serious consequences, including identity theft and insurance fraud. De-identification, or data anonymization, serves to hide an individual in the data, preventing the ability to single out a specific person.

This paper, the first in a series that explores de-identification and de-mystifies the techniques used to protect patient privacy, will cover the benefits derived from secondary data uses, discuss the risks organizations face from intentional or inadvertent data breaches, review existing legislation in the protection of PHI and introduce options organizations can use to de-identify data.

The Benefits of Secondary Use

Secondary uses of health data fall outside of delivering direct patient care. This covers activities such as research, analysis, monetization, measuring the quality and safety

The Benefits of Secondary Use

Recognizing the need for better health outcomes

Recognizing the need for better health outcomes, the State of Louisiana has been a bellwether in the provision of real-world data to foster healthcare innovation.

After placing 49th overall in a national health ranking, the State's Department of Health and Hospitals decided to launch an open data competition as a way to encourage budding entrepreneurs to create technology applications that would engage patients and empower them to make meaningful health decisions.

The success of the Cajun Code Fest resulted in the creation of 15 viable healthcare applications.

of the health system, epidemiological patterns, provider certification and accreditation, as well as marketing and other business applications.

Secondary use is becoming more common as organizations increasingly share their data holdings with third-parties. Leveraging this data provides opportunities to address challenging problems in healthcare, drive innovative research and gain new insights into patient care and therapeutic outcomes that can benefit the population as a whole.

Consent versus De-identification

In order for a person's PHI to be used for secondary purposes, current privacy laws require that the patient provide their consent for use. Once consent is obtained then the data can be used for the purposes that the person has authorized without the need to de-identify it. The trouble comes in trying to gain informed consent. Generally, at the time consent is being sought, all possible secondary uses of the data are unknown. Seeking consent at a future point, once specific secondary uses have been identified, may seem like a viable solution. However, contacting the thousands, or millions, of individuals affected can be an expensive, time-consuming and, ultimately, a futile activity. Some patients will have moved, some will have changed their contact information, and others will have died, making consent impossible to obtain.

De-identification is a best practice in data management when data is to be disclosed for secondary purposes; it allows PHI to be shared without the need to obtain user consent.

Considerations in the Sharing of Data

The requirements to de-identify data are tightly tied to the circumstances in which it will be shared – the nature of the data, the people it is being shared with and the goals of the analysis. When looking for solutions to de-identify datasets, Privacy Officers and data custodians must consider the trade-off between the need for a high level of privacy versus the need for data specificity. To achieve a low probability of re-identifying an individual, the information contained in the dataset will, by extension, be less useful for analysis. For example,

The Risks of Privacy Breach

if the data is to be made available to the general public, as it was in the State of Louisiana's Cajun Code Fest, a high degree of de-identification is required. This can hamper meaningful analysis but is necessary to minimize the risk to individual privacy.

In addition to de-identification, security controls and contractual obligations can also be used to support data protection. If data will be disclosed in a restricted access environment, then data sharing agreements and the use of regular audits help to ensure that adequate security and privacy practices are being maintained.

The Risks of a Privacy Breach

Organizations are putting a greater focus on the need to protect their information holdings; healthcare organizations, in particular, are under pressure to ensure sufficient protection. There have been several high-profile incidents recently, the largest of which was uncovered in February 2015 at U.S-based health insurer Anthem, when it was revealed that the records of almost 80 million patients were stolen.[1] Unlike attacks on retail or financial organizations where stolen credit card numbers can be easily cancelled and replaced, patient medical records include information that isn't easily destroyed, such as birth dates or Social Security Numbers. This makes health information highly valuable on the black market.[2]

Legal, Financial and Reputational Risks of a Privacy Breach

The healthcare sector currently offers some of the best standards for privacy. Legislation has been passed in many jurisdictions that put measures in place to protect personal information. In the United States, health information is protected by the Health Insurance Portability and Accountability Act (HIPAA), which marked its 20th anniversary in 2016. The HIPAA Privacy Rule was designed to protect PHI by permitting only certain limited uses and disclosures of a person's health information covered by the Rule unless otherwise authorized by the individual subject of the information. Detailed information about the HIPAA Privacy Rule can be found on the website of the U.S. Department of Health and Human Services Office for Civil Rights (http://www.hhs.gov/ocr/office).

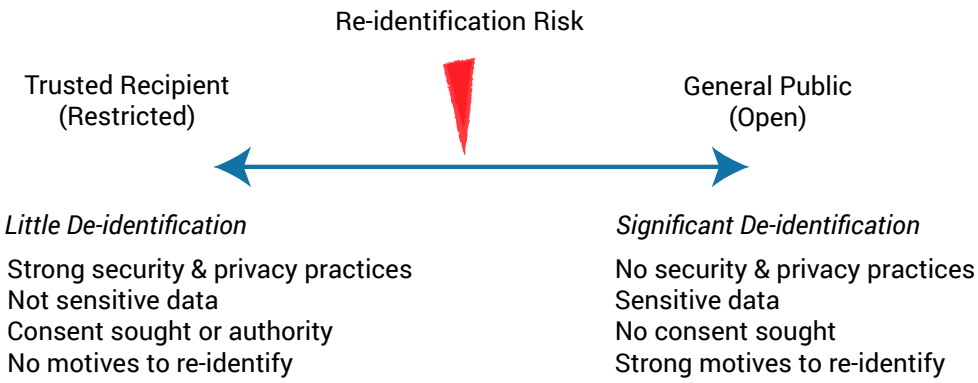


Figure 1: The tradeoffs in choosing the type of access to provide to a dataset

The Risks of Privacy Breach

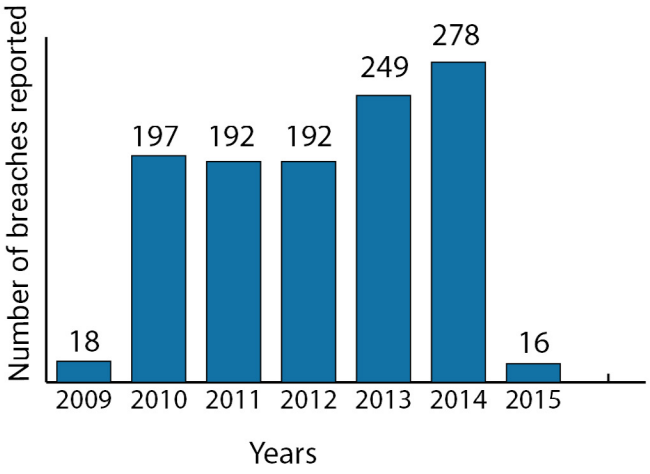


Figure 2: Number of Reported Data Breaches, as of Feb 24, 2015.

Legislation exists in Canada at both the national level through the Personal Information Protection and Electronic Documents Act (PIPEDA) and provincially, where some jurisdictions have adopted their own heath information legislation, such as the Canadian province of Ontario's Personal Health Information Protection Act (PHIPA). New The European Union recognizes the protection of personal information under the General Data Protection Regulation. Most of these laws include provisions that allow a person to file a complaint or take an organization to court if they believe that their data has been disclosed for an unauthorized purpose.

In addition to the legal implications, there are also financial and reputational consequences of a privacy breach. Legislation often mandates that organizations notify the people whose data was exposed.[3] Data breaches have been estimated to cost, on average, $208 per person to cover notification and legal action.[4] Furthermore, a breach can damage the reputation of an organization, erode public trust and impact the bottom line.

Intentional and Inadvertent Data Breach

According to the U.S. Department of Health and Human Services, large healthcare data breaches are on the rise. From 2009 to 2014, there has been a 15-fold increase in the number of large data breaches, where at least 500 people have been affected.[5]

Not only is the information stored in databases at risk from external threats like criminals and hackers, it is also subject to inadvertent data leaks from security vulnerabilities or employees. In March 2015, it was disclosed that Google had accidentally allowed the personal details of 280,000 domain name owners to be accessed as a result of a security vulnerability that permitted this information to be openly queried.[6] Unfortunately, simply removing information such as names and addresses from a dataset does not mean that the data is anonymous and that an individual cannot be identified. De-identified data from the Group Insurance Commission, which purchases health insurance for Massachusetts state employees, was matched against the voter list for Cambridge, allowing the Governor's record to be re-identified.[7]

Ensuring real privacy protection means taking steps to assess re-identification risk, set acceptable risk thresholds and apply de-identification standards to transform the data. De-identification is a key component of any risk management plan aimed at protecting individual privacy. While it is the goal of de-identification to

Data De-identification Options

reduce the chance of a person being re-identified to zero, it is impossible to guarantee this could never occur. Although it is very unlikely that you will be struck by lightning (1 in 960,000)[8], there are a small number of people that suffer this fate each year. However, like staying away from tall trees in a thunderstorm, de-identification helps us further mitigate the risk.

The HIPAA Privacy Rule provides mechanisms for using and disclosing health data responsibly without the need for patient consent. These mechanisms center on two HIPAA de-identification standards – Safe Harbor and the Expert Determination Method. Safe Harbor relies on the removal of specific patient identifiers (e.g. name, phone number, email address, etc.) while the Expert Determination Method requires knowledge and experience with generally accepted statistical and scientific principles and methods to render information not individually identifiable.[9]

Several options are available to organizations in how they go about de-identifying their data: they can use a homegrown solution, engage a de-identification consultant or purchase commercially available software tools.

Homegrown solutions tend to be based on an existing standard, such as HIPAA's Safe Harbor method, which does little to retain the analytic value of the dataset. HIPAA's Safe Harbor only looks at 18 different types of criteria, which has the potential to increase the risk of re-identification. Additionally, Safe Harbor was not conceived with longitudinal data – data collected over a given period of time – in mind, allowing even greater re-identification risk in these situations.

Engaging data de-identification experts who act as consultants and can certify that the data is de-identified each time it is to be shared can prove expensive and time-consuming. This solution also doesn't scale well; in cases where the dataset contains millions of records the process may become too lengthy to be workable. Most experts will not want to provide their methodology and this may leave the organization unable to prove that the approach used has produced a low risk of re-identification.

> De-identification is a key component of any risk management plan aimed at protecting individual privacy

Commercial software tools are now available that can perform automated in-house de-identification that applies HIPAA's Expert Determination method. These can provide a cost-effective and repeatable solution that use peer-reviewed techniques to measure re-identification risk and comprehensively de-identify data while retaining its analytic value.

Conclusion

Enabling personal health information for secondary use is critical to driving innovation, deriving insights and gaining new knowledge.

Conclusion

Contact Us

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

Phone: 613.369.4313

www.privacy-analytics.com

sales@privacy-analytics.com

Copyright@ 2017 Privacy Analytics Inc.

All Rights Reserved

When sharing information sources, organizations need to consider how they will balance protecting the privacy of the individuals whose information is contained within the data against retaining data quality for analytic purposes. The risks of privacy breach are real and carry with them serious legal, financial and reputational consequences. Data de-identification offers an effective way to mitigate the chances that an individual will be re-identified from the data while preserving the data's value. Homegrown solutions and de-identification consultants are options for implementing de-identification; however, commercial software tools can provide data custodians and Privacy Officers with a comprehensive and cost-effective data management solution.

Continue the expert's journey with the next paper in this series, De-identification 201: Fundamentals of Data De-identification.

This series is also available in webinar format: start by watching De-identification 101, available at www.privacy-analytics.com/de-id-university/webinars/.

Appendix: Terminology

Term	Definition
Anonymization	Sometimes used interchangeably with the term de-identification. A process that removes or suppresses, and/or alters personally identifiable information in a data collection so that it may be shared within the organization, with other organizations, or individuals for secondary purposes.
Business Associate	Business Associates are defined under HIPAA as a person or entity that performs certain functions or activities that involve the use or disclosure of protected health information on behalf of, or provides services to, a Covered Entity.
Covered Entity	Covered entities are defined under HIPAA as health plans, healthcare clearinghouses and healthcare providers that electronically transmit any health information. By law, the HIPAA Privacy Rule applies only to Covered Entities.
Dataset	A collection of related data records. Most commonly, a dataset refers to the contents of a database with many tables of data, where every column in the table represents a particular variable.
De-identification	See Anonymization. The term de-identification is used more frequently in the United States.
Expert Determination	Also referred to as Statistical Method. A standard methodology for de-identification specified under the HIPAA Privacy Rule. Expert Determination requires a person with appropriate knowledge of, and experience with, generally accepted statistical and scientific principles and methods to certify and document that a dataset is sufficiently de-identified such that there is a very small risk that an individual can be identified from the data.
HIPAA	The Health Insurance Portability and Accountability Act. Federal legislation enacted in the United States in 1996 that protects the confidentiality and security of personal healthcare information by setting limits on the use and disclosure of a person's data unless consent for additional secondary purposes has been obtained from the individual subject of the information (or is compelled by court order).
HIPAA Privacy Rule	Sets out the standard for privacy of individually identifiable health information. Contained within the Health Insurance Portability and Accountability Act, the HIPAA Privacy Rule applies to organizations that are defined as Covered Entities under the Act and requires that those that work with HIPAA Business Associates produce a contract that imposes safeguards on the PHI that the Business Associate uses or discloses.
Protected Health Information (PHI)	Health data that can be used to uniquely identify or locate an individual. Examples of protectedl health information include health plan numbers, disease diagnoses, hospital admissions information or lab results.
Personally Identifiable Information (PII)	Data that can be used to uniquely identify or locate an individual. Examples of personally identifiable information include name, phone number or credit card number.
PHIPA	Personal Health Information Protection Act. Ontario provincial legislation established in 2004 that provides a set of rules for the collection, use and disclosure of personal health information, requiring an individual's consent for these activities and requiring health data custodians treat all PHI as confidential and maintain its security.
PIPEDA	Personal Information Protection and Electronic Documents Act. Federal legislation enacted in Canada in 2000 primarily to support and promote electronic commerce. It governs how private-sector organizations collect, use and disclose personal information in the course of conducting commercial business.

Appendix: Terminology

Term	Definition
Privacy Breach	The result of unauthorized access to, or collection, use and disclosure of personally identifiable information for one or more individuals whose information is contained in a dataset. Privacy breach may be inadvertent (security vulnerability) or intentional (hacker).
Re-identification	The identification of a unique individual within a dataset that was supposed to have been de-identified.
Safe Harbor	A standard methodology for de-identification specified under the HIPAA Privacy Rule. The Safe Harbor methodology requires the removal of 18 types of direct and quasi-identifiers from a dataset so that no actual residual information can be used to identify an individual.
Secondary Use	Any use of a dataset other than for the provision of direct patient care. Secondary uses of healthcare data include research, analysis, monetization, quality and safety, accreditation, policy setting, and marketing or other business applications.
Statistical Method	See Expert Determination.

Sources

1) Abelson, Reed & Julie Creswell (2015, February 6). Data Breach at Anthem May Forecast a Trend. New York Times. Retrieved from http://www.nytimes.com/2015/02/07/business/

2) Ibid.

3) U.S. Department of Health and Human Services. Breach Notification Rule. Retrieved from http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/

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

5) Large Data Breaches on the Rise (2015, March 2). Retrieved from http://healthitmhealth.com/patient-privacy-wake-largest-medical-data-breach/

6) Price, Rob (2015, March 13). Google accidentally leaked hundreds of thousands of customers' personal details and didn't notice for 2 years. Business Insider. Retrieved from http://www.businessinsider.com/

7) Barth-Jones, Dan (2012, August 12). The Debate Over 'Re-Identification' Of Health Information: What Do We Risk? Health Affairs Blog. Retrieved from http://healthaffairs.org/blog/2012/08/10/the-debate-over-re-identification-of-health-information-what-do-we-risk/

8) National Oceanic and Atmospheric Administration National Weather Service. How Dangerous is Lightning? National Weather Service Lightning Safety. Retrieved from http://www.lightningsafety.noaa.gov/odds.htm

9) U.S. Department of Health and Human Services. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Health Information Privacy. Retrieved from http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html