Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification

Gregory S. Nelson

ThotWave Technologies, Chapel Hill, NC

Abstract

Researchers, patients, clinicians, and other healthcare industry participants are forging new models for data sharing in hopes that the quantity, diversity, and analytic potential of health-related data for research and practice will yield new opportunities for innovation in basic and translational science. Whether we are talking about medical records (e.g., EHR, lab, notes), administrative information (claims and billing), social contacts (on-line activity), behavioral trackers (fitness or purchasing patterns), or about contextual (geographic, environmental) or demographic (genomics, proteomics) data, it is clear that as healthcare data proliferates, threats to security grow.

Beginning with a review of the major healthcare data breaches in our recent history, this paper highlights some of the lessons that can be gleaned from these incidents. We will talk about the practical implications of data sharing and how to ensure that only the right people will have the right access to the right level of data. To that end, we will not only explore the definitions of concepts like data privacy but also discuss, in detail, various methods that can be used to protect data – whether inside an organization or beyond its walls. In this discussion, we will cover the fundamental differences between encrypted data, "de-identified", "anonymous", and "coded" data, and the methods to implement each. We will summarize the landscape of maturity models that can be used to benchmark your organization's data privacy and protection of sensitive data.

INTRODUCTION 2

WHY HEALTHCARE DATA? 2
DATA BREACHES – LESSONS LEARNED 3
PROTECTION OF DATA 5
DATA SHARING AND DATA TRANSPARENCY 6

UNDERSTANDING DATA PRIVACY 8

MEASURING THE "IDENTIFIABILITY" OF DATA 10
DE-IDENTIFICATION VERSUS ANONYMIZATION 12
METHODS OF DE-IDENTIFICATION 12

DATA PROTECTION MATURITY 15

TECHNIQUES FOR ENSURING DATA PRIVACY FOR ANALYTICS 15
INFORMATION SECURITY AND PRIVACY MATURITY 18

SUMMARY 20

BIOGRAPHY 21
CONTACT INFORMATION 21

REFERENCES 22

Introduction

In the United States, we function in a highly interoperable world. We take for granted that we can go online to locate our friends (e.g., Facebook) and colleagues (e.g., LinkedIn), and access our most personal information (bank and investment account balances; healthcare lab results.) At the same time, we face threats nearly every day. Email scams can introduce vulnerabilities to our computers and networks. Hackers seek to steal our secrets, implant viruses, or cause chaos or damage. And there is the omnipresent risk of theft or loss due to negligence, improper controls, or just poor processes and technology for prevention and detection.

As of December 2014, the Identity Theft Resource Center (ITRC) has tracked over 1,190 data breaches in the healthcare sector from 2005 -2014. In total, this represents approximately 25% of all of the data breaches for this same time period. Since 2005, breaches in healthcare alone are up 300% and represent 42% of those reported in 2014. Note that while the numbers vary depending on the messanger, conservative estimates report that more than 120 million people have been compromised since 2009. (Services, 2015)

Some have even given 2015 the moniker "the year of the healthcare hack" (Peterson, 2015). In fact, just within the first three months of 2015 we saw a nearly 91 million healthcare records reported as "compromised" with the two largest companies being health insurers: Premera Blue Cross (Reuters, 2015) and Anthem Blue Cross (Riley, 2015). With Anthem, that one incident more than doubles the number of people affected by breaches in the health industry since the Department of Health and Human Services started publicly reporting the issue in 2009.

Why Healthcare Data?

At first guess, it may seem far more interesting and lucrative to steal social security numbers and bank account information as opposed to healthcare data. But the reality is that healthcare data is often an easier target. In fact, the banking industry (banking/ credit/ financial services) has a 10-year average of experiencing data breaches of only 8.1% percent of all reported breaches and has reported the least number of breaches for nine of the past 10 years as compared to all other industries (Itrc, 2015). Healthcare, on the other hand, has experienced a tremendous rise in events (over 300% increase since 2005), which may be partly explained by the widespread adoption of electronic health records. On the black market, healthcare data can be used to file false tax returns, open lines of credit, or claim medical benefits. In fact, one of the most common forms of fraud is in acquiring prescriptions based on the identity of a patient (Wartell & La Vigne, 2015) making "Fraud, Waste and Abuse" one of the top priorities for health insurers as they seek to tap analytic talent.

Data Breaches – Lessons Learned

The following table highlights only a few of the cases in the past few years along with the number of lives affected and type of data lost.

Date Reported	Organization	Summary of Impact
March 2015	Anthem Health (Blue Cross of CA)	• Nation's second largest health insurer • Hackers accessed 80 million current and former Anthem members information (names, birthdays, medical IDs/social security numbers, street addresses, email addresses and employment information, including income) • Estimated costs for fixing problem--$100-$200 million not including penalties
August 2014	Community Health Systems (Tennessee)	• Operate 206 hospitals in 29 different states; #2 publicly traded hospital operator • Second largest breach ever • 4.5 million records (non-medical information; however, patient names, addresses, birthdates, phone numbers and Social Security numbers were obtained) • Summary: Chinese operatives hacked through a test server that was never intended to be connected to the internet; therefore, security features were not deployed. VPN credentials were stored on the test server, so when it was connected to the internet, hackers were able to access the test server via the Heartbleed bug and obtain the VPN credentials
February 2014	St. Joseph Health System (California)	• Not for profit, Catholic IDN, hospitals in California and Texas • 405,000 employee and patient records (for employees--banking information, names, addresses, dates of birth and Social Security numbers; for patients--some medical records and lab tests) • Summary: A hacker conducted a 3-day attack in December 2013 against an onsite server affecting 5 of the hospitals in Texas
March 2014	Sutherland Healthcare Solutions	• Business Associate/Vendor handling billing and collections for LA County's Department of Health Services and Department of Public Health • 342,000 records (patient names, Social Security numbers, billing information; may have also contained addresses and diagnoses) • Summary: Torrance, CA, office was broken into and 8 unencrypted desktops were stolen
September 2014	Touchstone Medical Imaging (Tennessee) Nationwide diagnostic imaging company	• 307,000 records (patient names, dates of birth, addresses, phone numbers, health insurance information, radiology procedures, diagnoses, etc.) • Summary: Folder with information from radiology procedures was "inadvertently" left accessible via the internet; billing information made temporarily available
October 2014	California Pacific Medical Center Breach	• An employee viewed improperly 844 patients; was terminated

Date Reported	Organization	Summary of Impact
		• Diagnoses, notes, demographics, prescriptions were compromised
August 2014	Indian Health (Maryland)	• 214,000 records • Summary: Employed physician improperly accessed information from patients at 3 IHS facilities

Table 1: Recent Healthcare Data Breaches

With the exception of Anthem and Premera, most of the data loss cases do not result from "hacking." Rather it is due to the negligence of individuals who use that data or are responsible for its well-being. Given this, it is important to recognize the factors that increase risk. There are valid business reasons for sharing data outside the boundaries of internal systems. One of the basic tenets of the federal government's Affordable Care Act (ACA) has been a desire to share data across entities in ways that support biosurveillance and improved continuity of care—enabling access to data for providers across all care environments as well as to the patients themselves. (Francis, 2015)

At the same time, there is increased imperative for population health management, which relies upon unaffiliated providers accessing data through health exchanges. However, we are faced with regulatory requirements that often lag years behind the technology and processes designed to protect that data. Furthermore, regulators do little to clarify how providers are to conduct risk analyses and which methodologies should be used, leaving us with a conflict between making data accessible and keeping it private and secure.

We also find that organizations struggle with finding developers who understand both the technical and administrative sides of healthcare security and clinically relevant data. The recent mandate to achieve meaningful use with electronic health records (EHRs) promotes greater interoperability between organizations, but most are under-resourced to deal with the magnitude and volume of data they handle as budgets are squeezed by competing IT priorities. Combined with the uncertainty of regulations and the pure costs of implementing stronger, safer solutions (note: most healthcare providers spend only about 3% of their IT budget on security (Conn, 2015)), we find that most organizations are ill-equipped to deal with this challenge. In fact, when recently helping to remediate an informatics platform for a leading edge, integrated delivery network in the United States, we still see examples of the wrong people having access to patient-level data and the storing of passwords in clear text on computers. On the consumer side, most patients are not educated as to their responsibilities in keeping data secure despite the growing number of access points to be protected in personal health information (e.g. mobile devices, browsers).

In contrasting these macro-environmental constraints with the rise in sophisticated methods of hacking given the value of personally identifiable information on the black market, it is no wonder that healthcare sits at a critical juncture. On one hand, we are moving toward an environment of massive data sharing capabilities with tremendous potential for the advancement of individual and population outcomes. On the other hand, we see evidence that our interconnectedness is exposing nearly a third of the US population to identity theft.

Protection of Data

It is important to consider the audience and its purpose as the sensitivity of the data and the analytics performed on that data are increased. Elements of the physical and information security protocols within an organization control access to the data. However, as you move further away your own organization, you must consider the sensitivity of that data, the risks of exposure and the measures you use to protect that data. The graphic here characterizes the core belief system of many organizations and their strategy for protecting sensitive data.

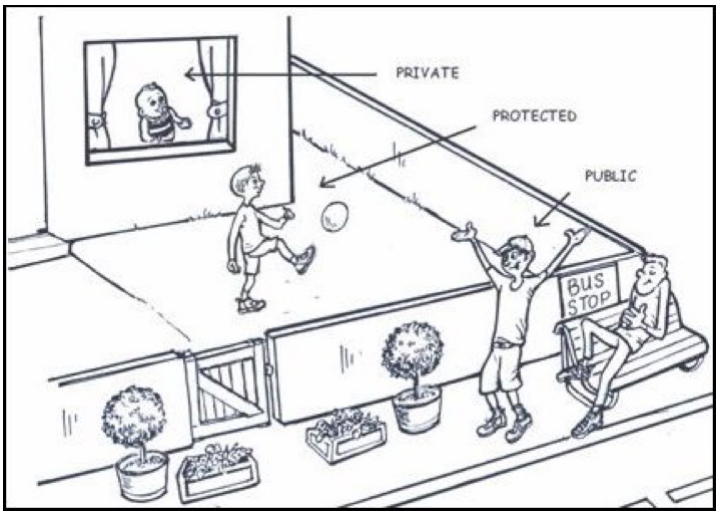


Figure 1: Simplistic model of Private, Protected and Pubic data (Source: Unknown)**.**

The model of protected data that lives within the firewall of most organizations is no longer sufficient to ensure that data is private, protected, secure, and usable. Nor does it take into account the "trusted relationships" that exist between the "household" – the modern organizations – and the rest of the world in which we operate.

Data Sharing and Data Transparency

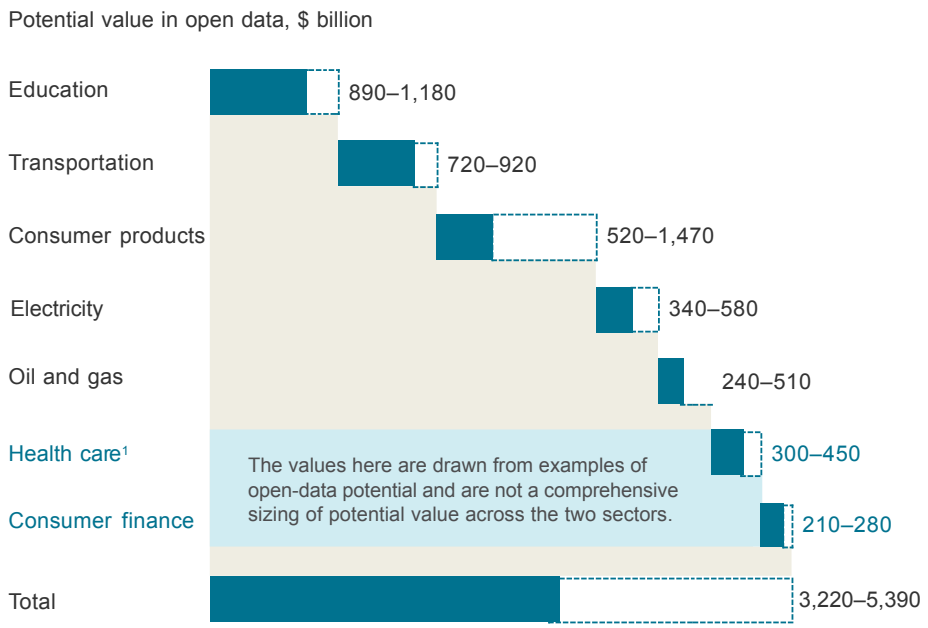


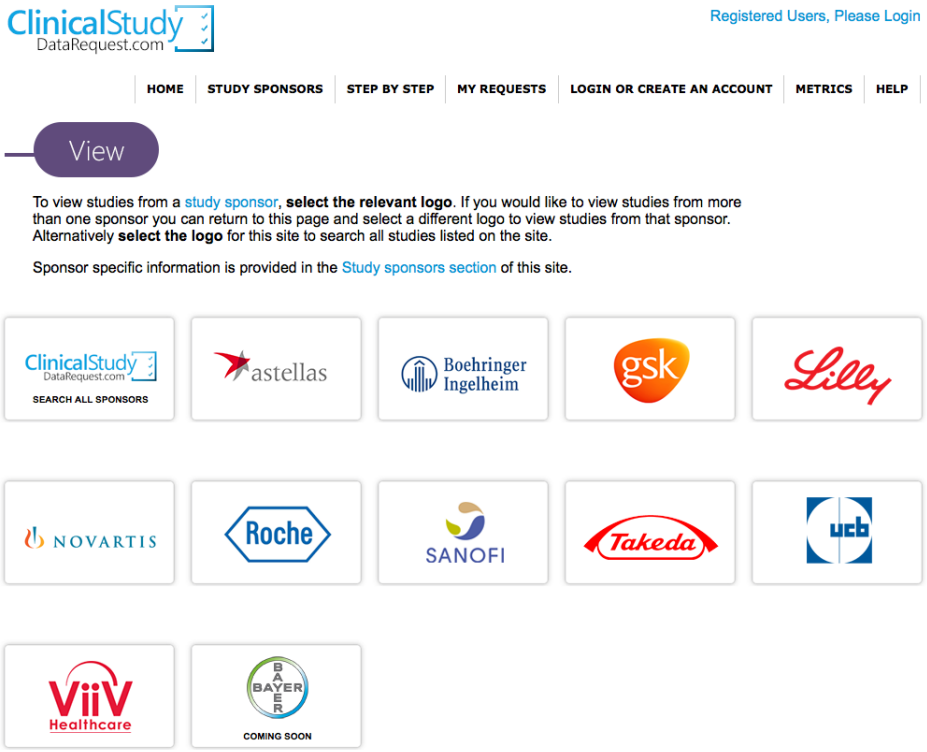
Figure 2: Potential Value of Open Data Across Industry (Source: McKinsey Global Institute).

Despite the need to keep data secure, there are tremendous advancements to be made in leveraging shared data intelligently for the improvement of community well-being. McKinsey (Manyika et al., 2013) estimates that the potential impact of data sharing in healthcare could amount to $300-$450 billion in annual value. They outline the role of "five levers" which help capture the value through the use of open and proprietary data.

As patients, we reap the benefits of our healthcare providers having access to complete records through the reduction of duplicate testing and improved access to our complete health records – even while we are traveling. Other potential benefits include:

- Enabling patients to take an active role in disease prevention and treatment
- Helping providers determine what is the most timely, appropriate treatment for each patient
- Matching patients with the most appropriate providers
- Ensuring the cost-effectiveness of care, and identifying new therapies and approaches to delivering care.

Most of the value comes in the form of cost savings to providers, payers, and patients. In the clinical trials world, there is a tremendous push towards transparency and access to data for "secondary use," that is, for additional research outside of the scope of original collection intentions. In 2013, GlaxoSmithKline announced a system that would allow researchers from around the globe access to anonymized patient-level clinical trials data (Nisen & Rockhold, 2013). Through a web site (https://www.clinicalstudydatarequest.com) other companies have followed suit and include data from a number of Biotechnology and Pharmaceutical companies as shown below.



The push to "transparency" in clinical trials comes with controversy. For example, the European Medicines Agency has recently revised its stance on data sharing for clinical trials as the opinions around controls, formats, and the sensitivity of data varied widely between companies and regulators. (Bonini, Eichler, Wathion, & Rasi, 2014)

As companies seek to improve their reputation and allow researchers to continue to find interesting patterns at the patient-level, it is important to define what open access to data means and how we ensure the privacy of individuals. The figure below summarizes McKinsey's perspective (Manyika et al., 2013) on the characteristics that "open data sets" share:

- **Accessibility:** A wide range of users is permitted to access the data.
- **Machine readability:** The data can be processed automatically.
- **Cost:** Data can be accessed free or at negligible cost.
- **Rights:** Limitations on the use, transformation, and distribution of data are minimal.

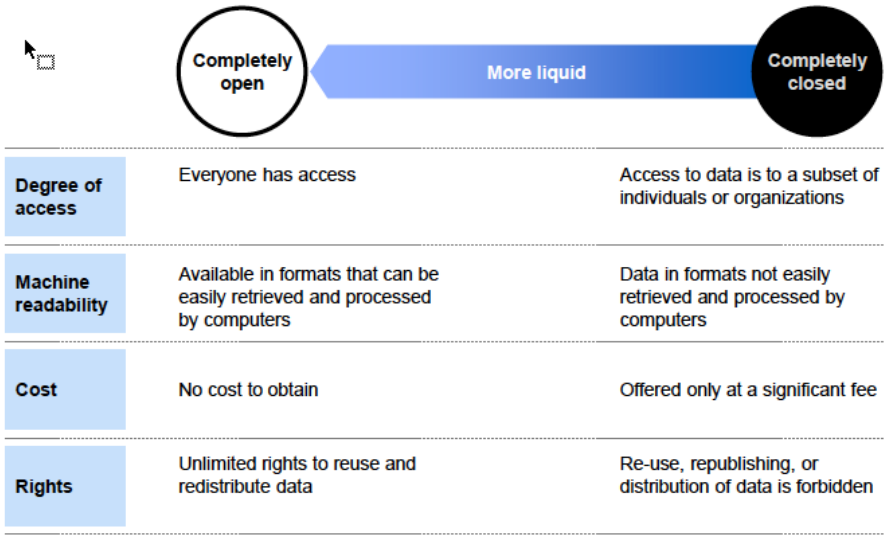


Figure 3: Characteristics of open or closed data (Source: McKinsey Global Institute).

Understanding Data Privacy

In the United States, one of the primary standards used to provide guidance for de-identifying personally identifiable information (PII) and personal health information (PHI) is the HIPAA Privacy Rule (45 CFR 164.514) from the US Department of Health and Human Services. It is designed to protect personally identifiable health information by either outlining appropriate uses and disclosures of PHI, or as authorized by the individual subject of the information. The HIPAA Privacy Rule also provides mechanisms for using and disclosing health data responsibly without the need for explicit patient consent. These mechanisms center on two HIPAA de-identification standards: HIPAA Safe Harbor and the Statistical or Expert Determination methods. (Rights, 2013)

The former standard specifies 18 data elements that must be removed or generalized in a data set in order for it to be considered "de- identified." The HIPAA Safe Harbor data elements (aka direct identifiers) include the following:

1. Names
2. Zip codes (except first three)
3. All elements of dates (except year)
4. Telephone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social security numbers
8. Medical record numbers
9. Health plan beneficiary numbers

10. Account numbers
11. Certificate or license Numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic or code

Table 2: Safe Harbor Data Elements

In the case of *Expert Determination* or *Statistical Method*, this first requires finding a person (or persons) who has appropriate experience with the rules governing identifiable information. They are fluent with the statistical methods and scientific principles for adjudicating the risk of data in terms of individual identification potential.

While the Safe Harbor method provides a simple way to meet the requirements, it is often criticized for not protecting the data from advanced methods of re-identification. In fact, some fairly publicized examples exist which illustrate this point:

- In 1997, using a known birth date, gender and zip code, a computer expert was able to identify the records of Gov. William Weld from an allegedly anonymous database of Massachusetts state employee health insurance claims (Barth-Jones, 2015).
- In 2007, with information available on the Internet, Texas researchers utilizing a de-anonymization methodology were able to re-identify individual customers from a database of 500,000 Netflix subscribers (Narayanan, 2008).
- In 2013, Science (Gymrek, McGuire, Golan, Halperin, & Erlich, 2013) reported the successful efforts of researchers in identifying "deindentified" male genomes through correlations with commercial genealogy databases.
- Students were able to re-identify a significant percentage of individuals in the Chicago homicide database by linking with the social security death index (K. El Emam & Dankar, 2008).
- AOL put de-identified/anonymized Internet search data (including health-related searches) on its web site. New York Times reporters were able to re-identify an individual from her search records within a few days (Porter, 2008).

For clinical trials conducted in the United States, research is governed by 45 CFR 46 (The US Code of Federal Regulations including the HIPAA Privacy Rule at 45CFR Parts 160 and 164, 2006), also known as

the "Common Rule", and applies to "all research involving human subjects conducted, supported, or otherwise subject to regulation by any federal department or agency which takes appropriate administrative action to make the policy applicable to such research." Jack Shostak makes an interesting comparison of the Common Rule and the HIPAA Privacy Rule in his paper "De-Identification of Clinical Trials Data Demystified" (Shostak, 2006):

- *The HIPAA Privacy Rule allows you to create and keep (and protect) a key that maps your old subject identifiers to the new de-identified identifiers. The "Common Rule" states that the key must be destroyed for the data to be anonymous.*
- *The "Common Rule" allows for new de-identified subject identifiers to be derived from old subject identifiers but the HIPAA Privacy Rule prohibits this.*
- *The "Common Rule" states that anonymized data should not contain any identifiable private information, but the HIPAA Privacy Rule lists 18 more specific data types to be excluded. The "Common Rule" generally allows for ZIP codes and dates to remain in anonymized data.*

The "guiding principles" outlined with the Common Rule and HIPAA Privacy Rule still force us as analytics professionals to devise ways to protect data. The degree to which data should or could be obfuscated of course depends on the usage goal and the degree of trust for those that use that data.

Measuring the "Identifiability" of Data

In order to make effective use of data, we must have data that is sufficiently clean, organized, and accessible. In this context, an EHR is looked to as the foundation for consistent, detailed data. An organization's internal data warehouse strategy will help standardize the way that questions are asked and ensure that data is collected and stored appropriately for analytics. Evolving Health Information Exchanges (HIEs) help socialize EHR data beyond a provider's firewall and make it accessible.

We live in a world where security by obscurity is no longer sufficient. As an industry, our challenge becomes how to structure data in ways where we can derive meaningful information while protecting the data from fraud, unintended use, or misinterpretation (and resultant inappropriate action.) We know, for example, that if we simply strip off the variables of interest (

Table 2: Safe Harbor Data Elements) then the data becomes far less useful to us as researchers and advanced analytic techniques. Furthermore, it may actually lead us to believe that our data is safe and secure – truly anonymized. But as we saw from the re-identification examples above, we may not be able to obscure the data well enough and still render the analysis useful.

While there are a number of techniques that can safeguard data, we will address just a few here. But in doing so, we put these into perspective using a model developed by Khaled el Elmam of the Children's Hospital of Eastern Ontario (Emam, 2010). He states that instead of considering the "identifiability" of data as a binary option, we should consider identifiability as a spectrum. In this context, various datasets fall on different points of the spectrum depending on ease, cost, and risks

imposed by re-identification. In the model shown below, the discrete levels characterize specific stages that a data set would go through as it is increasingly de-identified.

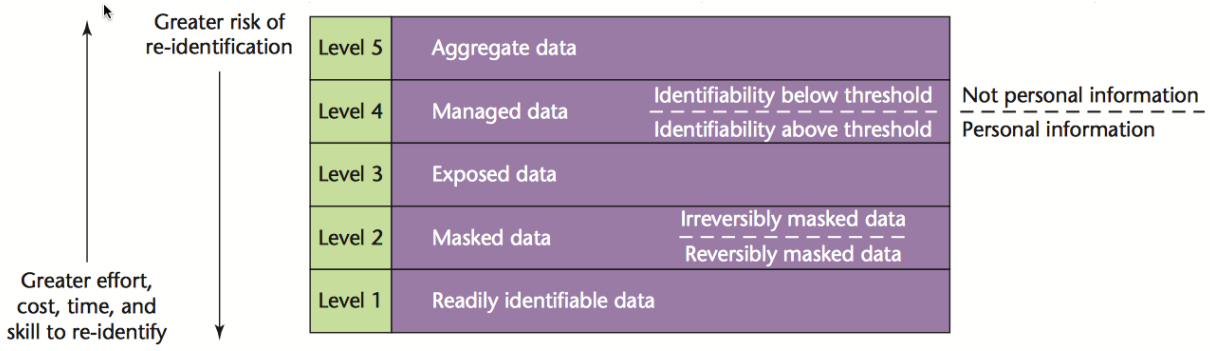


Figure 4: A conceptual five-level model of the identifiability continuum.

- **Level 1:** data that's clearly identifiable. For example, a database containing names, Social Security Numbers (SSNs), biometrics and dates of birth or other identifying information.
- **Level 2:** data that is masked or obscured. For example, you may modify the "identifying" variables through randomization and creating reversible or irreversible pseudonyms. Most BioPharma companies conducting clinical trials use this technique to mask treatment or intervention information.
- **Level 3:** Masked identifiers and non-identifiers. As with Level 2, the identifiers (such as name and date of birth) are masked, but with Level 3, we also mask variables that are considered to be quasi-identifiers. An example might be that we mask gender, but fail to mask the variable that contains the last date of a pap smear or pregnancy flag.
- **Level 4:** Managed data. This is where the researcher actively manages (and measures) the degree to which re-identification can occur. If the risk is low (according to an established benchmark) then the data is considered managed with or without personal information being considered identifiable.
- **Level 5:** Aggregate data that cannot physically identify individuals. Through the use of aggregation methods, non-stratified counts, frequencies or rates are shared. Note that not all aggregated data meet this requirement if the cell size for a given crossing of some combination of variables can lead someone to identify a particular individual. An example might be people who responded to a survey where the sample size of one particular variable – say race – is small enough to deduce who that individual might be (with or without additional data.)

De-identification versus Anonymization

When considering the fortification of personally identifiable data, we have a continuum of choices that help us understand how we can protect it. On one end, the data is completely identifiable and on the other we have data cannot be re-identified. Given the former, we must employ physical security and access controls to limit access and on the latter, we use controls over the dissemination of and the level of identifiability in the data. In the "in between" states, we have to worry about the management of the data (Level 2 through 4 using the "identifiability continuum" described above.) While the definition of the terms anonymization and de-identification may be evident, it is worth exploring these concepts in detail.

De-identification of data refers to the process of removing or obscuring any personally identifiable information from individual records in a way that *minimizes* the risk of unintended disclosure of the identity of individuals and information about them. **Anonymization** of data refers to the process of data de-identification that produces data where individual records cannot be linked back to an original as they do not include the required translation variables to do so.

Specific steps and methods used to de-identify information may vary depending on the circumstances, but should be appropriate to protect the confidentiality of the individuals. While it may not be possible to remove the disclosure risk completely, de-identification is considered successful when there is no reasonable basis to believe that the remaining information in the records can be used to identify an individual.

One approach is to simply replace personal information with surrogates that can later be used to look up the real values (for example, create surrogate keys as found with randomization or blinded studies). Alternatively, we can simply drop the columns (verbatim descriptions used in clinical trials adverse event reports) or recode the variables (age or age range instead of date of birth). In the following section, we detail some of the techniques that can be used to control the identifiability of data.

Methods of De-Identification

When discussing de-identification, we make a distinction between two types of variables: direct identifiers and quasi-identifiers (also known as indirect identifiers). Direct identifiers are fields that can uniquely identify individuals, such as names, SSNs and email addresses. In healthcare, we are referring primarily to the PII as defined by the HIPAA Privacy Rule (see

Table 2: Safe Harbor Data Elements). Direct identifiers are often not used in any data and statistical analyses that are run on the healthcare data. Quasi-identifiers are fields that can identify individuals but are also useful for data analysis. Examples of these include dates, demographic information (such as race and ethnicity), and socioeconomic variables. This distinction is important because the techniques used to protect the variables will depend on how they are classified.

As stated above, de-identification refers to the removal of fields as stored in systems so we minimize the risk that identifiable information can be used to reconstruct an individual's record. De-identification methods include:

Method	Description	Example
Record Suppression	• Removing data (e.g., from a cell or a row) to prevent the identification of individuals in small groups or those with unique characteristics • When the combination of quasi-identifiers (e.g., sex, race, zip code, diagnosis) presents too high a risk of re-identification to be released • Often used in public health reporting, geo-spatial analytics or secondary use datasets • This method may result in the loss of fidelity for small sub-groups • Usually requires additional suppression of non-sensitive data to ensure adequate protection of PII (e.g., complementary suppression of one or more non-sensitive cells in a table so that the values of the suppressed cells may not be calculated by subtracting the reported values from the row and column totals)	Drop the observations for those patients where the number of patients for any combination of zip code, age category and diagnosis code is below a given threshold (e.g., 5 people)
Cell Suppression	• Suppressing or masking the value of a single field	A field in a patient record containing a very rare disease
Randomization	• Retains the direct identifiers (name, phone number), but replaces their values with simulated (random) values • Reduces the probability of reverse identification • Often used in creating data sets for software testing where all fields must be present and have realistic looking values	Algorithm which randomly replaces the date of birth for patients
Shuffling	• Data for one or more variables are switched with another record • Often used in creating data sets for software testing where all fields must be present and have realistic looking values • All of the values in the data set are real, but they are assigned to the wrong people	Distinct values of a variable are randomly assigned to records
Creating Pseudonyms or Surrogate	• The creation of aliases can be done in one of two ways where a variable such as SSN or medical record number is replaced with a surrogate • Refers to the unique descriptor that can be used to match individual-level records across de-identified data files from the same	Applying a one-way hash to the variable using a secret (protected) key. A hash is a function that converts a value to another value (the hash value) but you cannot reverse the hash value back to the original value

Method	Description	Example
	source (e.g., for the purposes of comparing health states over time) • Depending on the need, this can be done so that a key can be used to restore the original value or irreversibly (anonymized) • Has the advantage that it can be recreated accurately at a later point in time on a different data set	
Sub-Sampling	• Taking a random sample of a data set • Can also be taken using stratification to ensure that the proportion of class variables are the same as the original (e.g., age groups, gender, race)	Randomly select a sample (e.g., 10%) based on the original dataset size
Aggregation/ Generalization	• Rare quasi-identifiers can be aggregated to provide better de-identification or anonymization.	A low population postal code can be aggregated to a larger geographic area (such as a city). A rare medical profession, such as perinatologist, can be aggregated to a more general obstetrician
Adding Noise	• Often used to introduce noise or randomness in continuous variables • May have limited protection as there are methods used in signal processing techniques to remove the noise	Jittering can be used to add random noise to data (for example, to prevent overplotting in statistical graphs)
Character Scrambling	• Rearrangement of the order of the characters in a field • This has limited value as it may be quite easy to reverse and is not a reliable way to protect information	For example, “SMITH” may be scrambled to “TMHIS”
Character Masking	• Character masking is when the nth character or characters of a string are replaced with another character. • Simple methods that only replace the first or last character has limited use as the values can be reconstructed with little effort	Replace SMITH with SMIT* or *M*T*
Truncation	• A variant of character masking in that the nth character is removed rather than replaced with a special character	Replace SMITH with MITH or SMIT or SITH
Encoding	• The value is replaced with another meaningless value • Most effectively used when creating a surrogate value for unique values	Replace SMITH with X&T%#
Blurring	• Used to reduce the precision of the data	Convert a continuous variable into a categorical data elements, aggregating data across small groups of respondents, and reporting rounded values and ranges instead of exact counts

Method	Description	Example
		Replace an individual's actual reported value with the average group value (on more than one variable with different groupings for each variable)
Masking	• Used to "mask" the original values in a data set • The purpose of this technique is to retain the structure and functional usability of the data, while concealing information that could lead to the identification, either directly or indirectly, of an individual value	Replace sensitive information with realistic but fake data Modify original data values based on pre-determined masking rules (e.g., by applying a transformation algorithm).
Perturbation	• Involves making small changes to the data to prevent identification of individuals from unique or rare population groups • Data perturbation is a data masking technique in that it is used to "mask" the original values in a data set to avoid disclosure	Swap data among individual cells to introduce uncertainty, so that the consumer of the data does not know whether the real data values correspond to certain records, and introduce "noise," or errors (e.g., by randomly misclassifying values of a categorical variable)
Redaction	• The process of expunging sensitive data from the records prior to disclosure	All identifiers and quasi-identifiers are dropped from the dataset

Table 3: Types of De-Identifications Methods

It is important to keep in mind that even the masking techniques that are protective will significantly reduce the utility of the data. Therefore, masking should be applied only to the fields that will not be used in any data analysis, which are often the direct identifiers: fields such as names and email addresses that are not usually part of any analysis performed on the data. Also, one should not apply masking techniques to dates or geographic information because these fields are often used in data analysis, and masking would make it very difficult to perform an analysis using those fields.

De-identification is based on characteristics of the different variables and field type. For instance, different algorithms are applied to dates of birth than zip codes. A detailed discussion of the de-identification algorithms that we use can be found here - (K. El Emam et al., 2009). Because many data sets consist of both quasi-identifiers and direct identifiers, in practice it is important to apply both data protection techniques: masking and de-identification.

Data Protection Maturity

Techniques for Ensuring Data Privacy for Analytics

As analytics professionals it is our duty to ensure that the data that we access, integrate, analyze and disseminate carry the expectation of privacy access is controlled proactively. In our consulting practice, we often see data protection measures focus on the protection of data in source systems and in

centrally-managed data warehouses. However, a gap exists in many organizations for governing analytics sandboxes and development processes. Take for example, a typical process whereby a researcher has acquired access to data:

1. Attend HIPAA or Security training
2. Request access to patient-level data
3. DBA or Security Officer grants access to data
4. The system is audited to ensure that non-authorized users can query against the data

While simplistic, this model falls short of truly protecting the sensitive nature of that data. Examples of the unintended breach can include:

- Motivated programmer creates a map using the Google API to show doctors what can be done by positioning patient information on a map (disclosing patient locations for low incidence zip codes).
- Analyst uses visual data discovery tool on her (unencrypted) laptop to depict interesting patterns of patient encounters (potential loss of laptop and exposing of patient records).
- Statistician develops a readmission rate model and sends to a colleague to help verify the model through email with the hold-back dataset attached (non-sure email is subject to network sniffing, email hacking or lost portable storage).
- Claims analyst links billing information with patient records in Excel (non-sure storage of patient information).
- Developer designs a system for fraud detection and the business logic for matching encounters with billing data is stored in the version control system (while the original file system may be protected through ACLs, other developers have access to the version control system).
- Programmer creates ED Volume Report on the development server where other report writers are doing their work (Non-AES secure credentials are subject to replay attacks).
- Researcher diligently backs up his work using the companies' backup software (the policies that govern the system administrator's password management processes have no provision for protected data assets).

These types of activities occur each and every day, yet most security programs do not adequately account for the myriad activities that programmers, statisticians, econometricians, forecasters and other analytics professionals perform on a daily basis. We argue that most security tactics govern IT processes, but fail to cover analytic practices. Below, we have summarized some of the things that can be done to help improve security and protect data used internally and for data sharing/ data transparency initiatives.

Method	Description
Physical Protection	• Controlled access to physical storage (computers, storage, networks, backup systems, thumb drives, portable computing, access to research systems) • For virtualized machines, controlled access to images and limited permission to import/ export information including through the clipboard • Governance around backup, mirroring, and version control systems
Encryption or Cryptography	• Refers to the conversion of data into scrambled code before it is transmitted over a public or private network using an encryption algorithm • The encryption algorithm uses a string of bits known as a "key" to perform the calculations needed to transform the data into a new form that cannot be read without the use of that key • For protecting data in transition, encryption is critical for ensuring confidentiality, integrity and in some cases even authenticity • Avoid the use of non-secure protocols for network transmission of clear text; use strong encryption to store the credentials in a centralized location and only transmits the credentials using secure data transmission protocols.
Password Management	• Limited access to system or service accounts. • Full audit trail using individual accounts for performing service functions (e.g., sudo versus direct root access). • Use of password vault and/or rotating passwords (e.g., CyberArk, RSA Tokens) • Fully encrypted passwords using AES or Kerberos rather than simple encryption that could be subject to replay attacks • Physical and application level / metadata controls to ensure that physical security is not bypassed when accessing data
Protected Data	• Limit the use of shared credentials for accessing databases • Encryption file system and file contents • Physical files protected via passwords (SAS datasets, SAS catalogs, PDF files, Enterprise Guide Projects and Metadata bound libraries) • Proactively manage the de-identification of shared data, including measurement and monitoring the risks of internal and external analytic processes

Table 4: Types of Data Security Protection

During the recent "hacking" of Sony systems, the hack occurred through securing the credentials to a database that housed sensitive data. In this scenario, a user could gain access to an administrator's password either by guessing (usually based on a computer program that generates guesses) or through direct access (e.g., weak passwords, passwords stored or transmitted in clear text).

Often times we perceive our systems as being "secure" because we have standard operating procedures and a defensive cry of "we use Kerberos," but this may actually be creating a false sense of security as even encoded passwords can be used in what are referred to as "replay attacks". That is, even if a hacker doesn't know the original password for a given user, if they can obtain the encoded

password, they can use that in their attempt to login to a service to access data and gain access rights as if the original owner.

There are a number of potential solutions to this problem such as disallowing cached passwords in client applications; forcing the user to enter a password each time; using biometric devices such as retinal or fingerprint scanners; deploying service-oriented applications which generate new passwords (e.g., RSA tokens); or using vault solutions which store privileged account credentials (e.g., CyberArk, WebPasswordSafe) so that they can be changed on a regular basis.

Information Security and Privacy Maturity

Throughout this paper we have detailed the challenges surrounding data privacy and ensuring that data is protected. On one hand, we want to provide analytics professionals with well-integrated, high quality data to drive insights and actionable results. On the other hand, we have seen that de-identification techniques often fall short of protecting the individual who consented to providing data. Whether we are talking about data from a clinical trial or human genome data captured in the course of diagnosis and intervention, any identifiable data must be protected.

Organizations that rely on outdated regulations or information security practices designed for systems of old must evaluate their maturity on a scale relative to how analytics are used in the modern enterprise.

Privacy and confidentiality concerns hamper access to data. The potential for using de-identified data and for analysis will, no doubt, lead to improved healthcare, faster and more comprehensive public health surveillance, longitudinal analysis, and improved quality evaluation.

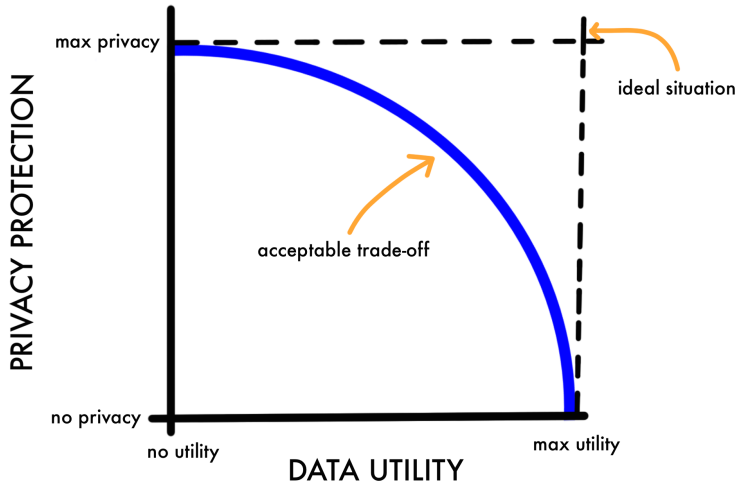


Figure 5: Data Privacy Protection versus Data Utility.

In the figure above, we see this tension at work but most organizations do little to actively manage the complexities that it implies.

Solutions such as the *De-Identification Maturity Model* (K. a. H. El Emam, Waël, 2015) help us characterize an organization's capability with respect to practices around preparing de-identified data. We believe this is a good start but attention should also be paid to other aspects of data privacy including physical protection, encryption of data and networks, and proper process and controls for access in addition to the protection of the data and its release.

In the figure below we illustrate the comparable risk present in the five levels of the model described above. We believe that the protection of the physical data and systems must be done so along with the de-identification strategies discussed in this paper. Even with level five de-identification strategies, you are never risk free as the techniques for re-identification evolve often faster than security measures intended to protect data. As for all of the levels, the physical protection and deep insight into the analytic life-cycle processes are paramount to helping preserve and protect data from unwanted prey.

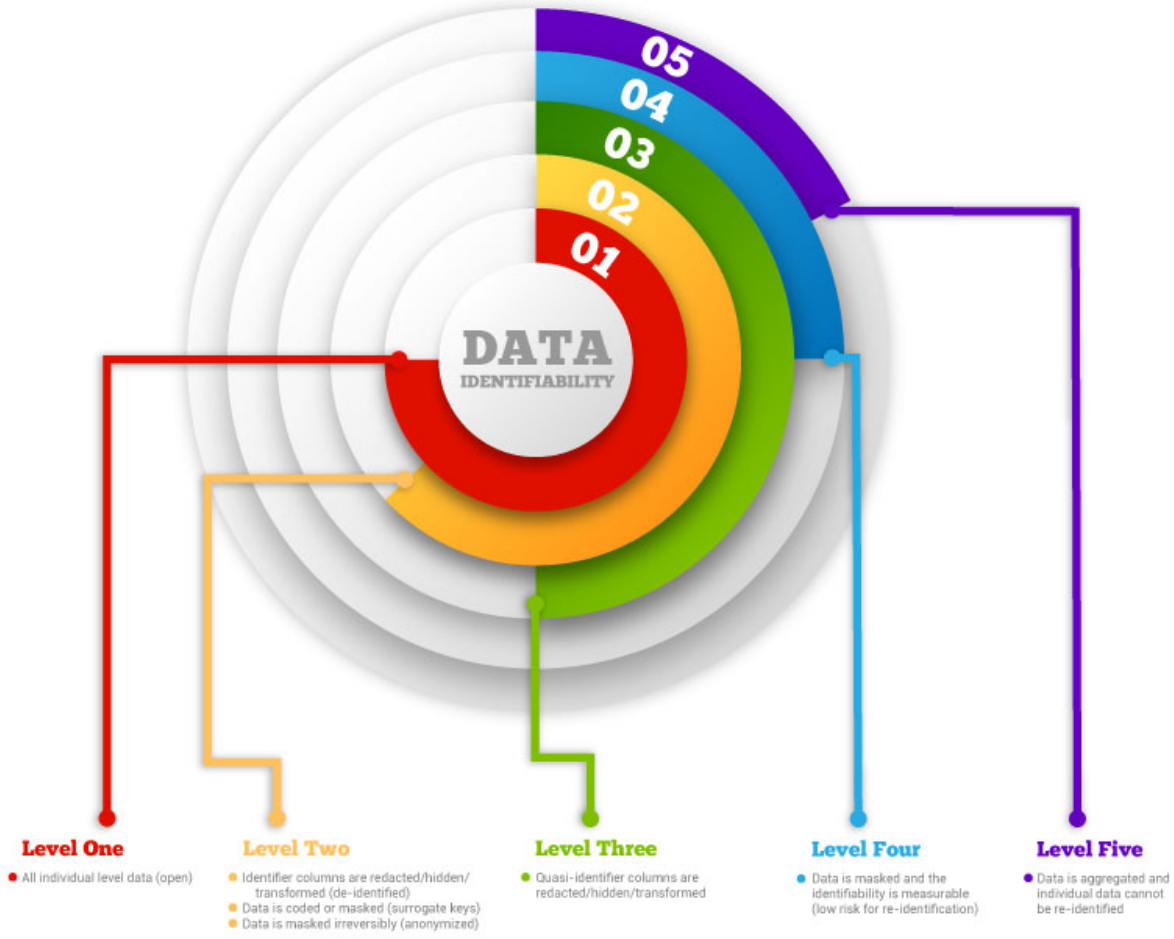


Figure 6: Data Identifiability Continuum.

In our consulting practice, we see little attention being paid to the ***analytic processes*** surrounding the protection and release of data, which should necessarily include the research/ insight life-cycle; development; testing and release processes; and metadata management.

Summary

We live in a world with a tremendous amount of electronic data floating around us - birthdates, social security numbers, zip codes, gender, marital status, hair color, prescription drug history - not to mention the droves of information that we voluntarily disclose via social media. While we certainly face the ethical challenges of how we as analytics professionals deal with onslaught of data (van Wel & Royakkers, 2015), we have a tremendous responsibility to ensure that the databases that we use every day, the analytics that we perform, and the tools that we use to access data are protected.

In this paper, we have highlighted some of the reasons why healthcare data is particularly important to protect and are witnessing the massive number of lives affected when data gets into the wrong hands. We also discussed the value of data sharing and the potential impact on the economy, not to mention the potential to improve the diagnosis, treatment and outcomes for patients. As we have noted elsewhere (Nelson, 2009), there has been a convergence across the health and life sciences space such that advances in translational medicine are creating therapies that can directly impact a patient on an expedited timeline. By opening up patient-level data to researchers around the globe, we hope that the translation between basic and applied research can help as BioPharma companies release their data reservoirs for all to partake.

We are guided by regulatory guidelines such as the HIPAA Privacy Rule and the Common Rule, but as we have noted, there remains the potential to re-identify data if mature processes are not followed.

As a profession, we need to do everything we can to ensure that the right data is getting to the right people in the right format. Technologies have evolved to the point where we can certainly manage this in a more mature manner. It is our responsibility to ensure that we do so promptly.

Acknowledgements

I would like to thank Monica Horvath and MaryLu Giver for their "thot-ful" review of this manuscript and especially to Carol Sanders for her research. Their combined insights and support of the healthcare analytics work that we do continues to inspire me.

Biography

Greg Nelson, President and CEO, Thotwave Technologies, LLC.

Greg is a global healthcare and Business Intelligence (B.I.) executive with over two decades of experience and leadership in the field. Greg is a prolific writer and speaker interested in healthcare analytics and the strategic use of information technology.

He received his BA in Psychology from the University of California at Santa Cruz and advanced his studies toward a PhD in Social Psychology and Quantitative Methods at the University of Georgia. Recently, Greg completed his Masters degree from Duke University in Clinical Informatics from the Fuqua School of Business. His academic and professional interests include helping organizations mature their analytic capabilities. Founder, President, and CEO of ThotWave Technologies, a niche consultancy specializing in healthcare analytics, Greg is particularly interested in how lessons from across other industries can be applied to help solve the challenges in healthcare.

With certifications in Healthcare IT, Project Management, Six Sigma and Balanced Scorecard, Greg is also a prolific writer and has presented over 200 professional and academic papers in the United States and Europe. He won the Best Paper Award in 2013 at the Pharmaceutical SAS Users Group Conference and sits on the board of the SAS Global Users Group. In 2011, Greg was selected by SAS into their loyalty partner group. "This program acknowledges and supports individuals who are recognized experts in their fields and have a long-term relationship with SAS."

Married to wife Susan and living on a small "farmlet" in rural North Carolina, Greg is an avid woodworker, enjoys photography, rides a Harley-Davidson Motorcycle, and strives to be a lifelong learner.

Contact information

Your comments and questions are valued and encouraged. Contact the authors at:

Greg Nelson greg@thotwave.com

ThotWave Technologies, LLC

1289 Fordham Boulevard #241

Chapel Hill, NC 27514 (800) 584 2819

References

Barth-Jones, D. C. (2015). *The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now*. Retrieved from http://papers.ssrn.com/abstract=2076397

http://papers.ssrn.com/sol3/Delivery.cfm?abstractid=2076397

Bonini, S., Eichler, H.-G., Wathion, N., & Rasi, G. (2014). Transparency and the European Medicines Agency — Sharing of Clinical Trial Data. *New England Journal of Medicine, 371*(26), 2452-2455. doi: doi:10.1056/NEJMp1409464

Conn, J. (2015). HIMSS survey: Healthcare providers boost security spending. from http://www.modernhealthcare.com/article/20140220/NEWS/302209951

El Emam, K., & Dankar, F. K. (2008). *Protecting Privacy Using k-Anonymity* (Vol. 15).

El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., . . . Bottomley, J. (2009). A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. *Journal of the American Medical Informatics Association : JAMIA, 16*(5), 670-682. doi: 10.1197/jamia.M3144

El Emam, K. a. H., Waël. (2015). The De-Identification Maturity Model. from http://www.privacyanalytics.ca/wp-content/uploads/2013/07/DMM.pdf

Emam, K. E. (2010). Risk-Based De-Identification of Health Data. *IEEE Security and Privacy, 8*, 64-67. doi: 1540-7993

Francis, B. (2015). The Affordable Care Act: How to Effectively Share Patient Data | Blog | Macadamian. from http://www.macadamian.com/2014/06/13/accountable-care-act-effectively-share-patient-data/

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying Personal Genomes by Surname Inference. *Science, 339*(6117), 321-324. doi: 10.1126/science.1229566

Itrc. (2015). 2014 Data Breaches | ITRC Surveys & Studies | ID Theft Blog. from http://www.idtheftcenter.org/ITRC-Surveys-Studies/2014databreaches.html

Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Almasi Doshi, E. (2013). Open data: Unlocking innovation and performance with liquid information. In M. G. Institute (Ed.).

Narayanan, A. a. S., V. (2008). *Robust De-anonymization of Large Sparse Datasets* Paper presented at the IEEE Symposium on Security and Privacy, 2008, Oakland, CA. http://arxiv.org/pdf/cs/0610105.pdf

Nelson, G. (2009). *Data Convergence in Life Sciences and Healthcare: Overview and Implications - 165-2009.pdf.* Paper presented at the SAS Global Forum 2009, Washington DC.

Nisen, P., & Rockhold, F. (2013). Access to Patient-Level Data from GlaxoSmithKline Clinical Trials. *New England Journal of Medicine, 369*(5), 475-478. doi: doi:10.1056/NEJMsr1302541

Peterson, A. (2015). 2015 is already the year of the health-care hack — and it's only going to get worse. Retrieved March 28, 2015, 2015, from http://www.washingtonpost.com/blogs/the-

switch/wp/2015/03/20/2015-is-already-the-year-of-the-health-care-hack-and-its-only-going-to-get-worse/

Porter, C. C. (2008). De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information - vol5_no1_art3.pdf. *Washington Journal of Law, Technology & Arts, 5*(1).

Reuters. (2015, March 17, 2015). Premera Blue Cross Says Data Breach Exposed Medical Data. *Business Day.*

Rights, O. f. C. (2013, 2013-07-26 00:00:00.0). Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. from http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

Riley, C. (2015, 2015-02-04T11:21:06). Insurance giant Anthem hit by massive data breach. from http://money.cnn.com/2015/02/04/technology/anthem-insurance-hack-data-security/index.html

Services, U. D. o. H. a. H. (2015). Breach Portal: Breaches Affecting 500 or More Individuals Retrieved March 28, 2015, 2015

Shostak, J. (2006). *De-Identification of Clinical Trials Data Demystified*. Paper presented at the PharmaSUG, Denver, CO.

The US Code of Federal Regulations including the HIPAA Privacy Rule at 45CFR Parts 160 and 164, the "Common Rule", 45CFR46 C.F.R. § Parts 160 and 164 (2006).

van Wel, L., & Royakkers, L. (2015). Ethical issues in web data mining. *Ethics and Information Technology, 6*, 129-140.

Wartell, J., & La Vigne, N. G. (2015). Center for Problem-Oriented Policing | Problem Guides | Prescription Fraud. 2nd Edition (2012). from http://www.popcenter.org/problems/prescription_fraud/print/