A
Synopsis
on
Analysis and Anonymization of Healthcare Data

Submitted in partial fulfillment of the requirements

for the award of the degree of

Bachelor of Technology

in
Computer Science and Engineering
by

Ashish Shukla, 1429010030
Mohit Kumar Sahni, 1429010078
Sourav Aggarwal, 1429010137

Under the Supervision of

Prof. Bipin Kumar Rai

ABESIT College of Engineering & Technology

Ghaziabad 201306

Affiliated to

Dr. APJ Abdul Kalam Technical University
Lucknow

Signature of guide

INDEX

Introduction

The amount of data is growing at an exponential rate, it's nearly doubling every two years and changing how we live in the world. One of the most important and critical data being the one produced by healthcare departments. As the confidential data accumulates further and further the risk of it's exposure becomes greater. Infact formal laws are being amended for protection of individual data. Notably GDPR , where General Data Protection Regulation (GDPR) (Regulation EU 2016/679) is a regulation by which the European Parliament, the Council of the European Union and the European Commission intends to strengthen and unify data protection for all individuals within the European Union (EU).

GDPR is set to be applied from 25 May 2018 in EU. So the question arises, *How do we ensure confidentiality of data ? In specific healthcare data.* Suddenly the conventional thoughts on Information Security evoke Encryption as the solution for every Information Security need. But to our surprise Encryption is not a solution or rather a Bad solution. The argument against Encryption is that it poses linear overhead of deciphering and enciphering information and being dependent on Keys the major drawback in this case is *what if the keys for data are leaked?*. And *Does enciphering data provides true confidentiality of our personal information?*
No.
As our identity is still attached to the data with the help of a key it can be directly seen to whom does the information belong. Thus although it satisfies basic criteria of GDPR but GDPR recommends a method called Pseudonymization.

Pseudonymization is a technique in which we replace quasi-specifiers with codes. Where quasi-specifiers are the fields in a dataset that identify the individual to whom the dataset belongs, it can be ZIP Code, DOB, Gender, Name, Aadhar Number etc. Partial Encryption maybe one way to visualize this technique.

In a scenario where the datasets must be fascilitated to various departments for further processing or analysis. If the data is not pseudonymized upon sharing the data it reveals identity of the individual. But pseudonymization hides the individual's identity and also removes the overhead of deciphering data for analysis or processing.

But the process of retrieiving identity of a patient from the pseudonymized dataset isn't really tough once we get the hold of the pseudonym table. Thus it is required to introduce concept of anonymization, where data anonymization is a type of information sanitization whose intent is privacy protection. It is the process of removing personally identifiable information from datasets completely so that re-identification is impossible.

Now the second problem is *Control over 'who should be able to identify data and who should not be.'* The problem can be solved using Role Based Access Control (RBAC). The control of who can access the data of an individual should be with individual himself.

In terms of Healthcare data A patient should be able to control who can access his records, e.g. A doctor who is examining a patient should be able to access his previous records. A family member upon being authorized by patient should be able to share his records.
The access can further be refined to which particular reports and fields an authorized person can access.

Brief Literature Survey

The usage of patient data for research poses risks concerning the patients' privacy and informational self-determination. Technologies like NGS(Next Generation Sequencing) and various other methods obtain data from biospecimen, both for translational research and personalized medicine. If these biospecimen are anonymized individual research can not be associated back to the individual. Thus we need some means to reassociate the data to individuals.
This technique is called de-identification and re-identification. Anonymization completely removes the re-identification possibility whereas pseudonymizatioin retains it. As the healthcare data is essential for Research purposes it is evident that pseudonymization is a better option than anonymization for it.

The consequences of Pseudonymization are that it does not provide complete assurance of confidentiality of identity. But it removes the threat of direct-identification.
As the clause 125 in IHE (Integrating the Healthcare Enterprise) Infrastructure Handbook states-
"*It is important to understand that you can only reduce the risks. The only way to absolutely assure a person can not be relinked to their data is to provide no data at all. De-identified data can still be full of identifying information. And may still need extensive privacy protections."*

So it is essential to optimize de-identification process to remove or in case of pseudonymization , pseudonymize identifying information completely or to an extent that it can not be relinked to the patient.

De-identification and Pseudonymization –

De-identification is the process of removing or transforming sufficient information from the source data. The goal is that the risk of re-identification is reduced to an acceptable level while also achieving the objectives of the intended use.
Pseudonymization is a particular type of de-identification that both removes the association with data subjects and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms.
In pseudonymization we may not necessarily replace each data with a unique pseudonym but we can also replace a batch of values for one pseudonym.
Data fields that are less identifying are usually not pseudonymized.

- **Types of Pseudonymization –** Based on reversibility there are two types of Pseudonymization -

1. Irreversible Pseudonymization – The pseudonymized data do not contain information that allows the re-establishment of the link between the pseudonymized data and the data subject. This overlaps with anonymization. But preserves continuity for the pseudonym throughout the resulting dataset.

2. Reversible Pseudonymization – The pseudonymized data can be linked with the data subject by applying procedures restricted to duly authorized users.

Our primary interest for re-identifiable data would be Reversible Pseudonymization.

Protection of healthcare demographics can be achieved using perturbative methods such as noise addition and data swapping. However these methods fail to preserve data truthfulness e.g. they may change the age of a patient from 50 to 10, which can severely harm the usefulness of the published patient data.
Non perturbative methods preserve data truthfulness.

It is important to remember data that is appropriately de-identified for one purpose may not be correctly de-identified for a new use of the data.

- **Algorithms for de-identification -**
  The major algorithms used in de-identification are -
  1. Redaction – Removing data or replacing it with missing data indicators.
  2. Fuzzing – Adding noise to data.
  3. Generalization – Making data less specific.
  4. Lognitudinal consistency – Modifying data so that data from many records remain consistent.
  5. Recoverable Substitution – Providing the ability to recover the original data values.
  6. Text Processing – Manual processing for text.
  7. Pass-through – Unmodified data is preserved in the resulting dataset.

Anonymization – Technically we use either one of the two techniques for anonymization -
1. Generalization – Replacing the values with a less specific but semantically consistent value.
Or
2. Suppression – Do not release a value at all.

- **Principles of anonymization -**

There are following principles of anonymization -

1. k-Anonymity – k-Anonymity is satisfied when each tuple in a table T(a1.....ad) where ai, i = 1...., m are Quasi-Identifiers(QIDs), is indistinguishable from at least k-1 other tules in T with respect to the set { a1,....., am} of QIDs.
K-Anonymity offers protection against identity disclosure, because the probability of linking an individual to their true record based on QIDs is no more than 1/k . The parameter k controls the level of offered privacy and is set by data publishers, usually to five in the context of patient demographics. We also note that not all the attributes in T need to be QIDs and that an individual may not be associated with some of these attributes.

The process of enforcing k-anonymity is called k-anonymization.
It can be performed by partitioning T into groups of atleast k tuples and then transforming the QID values in each group so that they become indistinguishable from one another. Note that an individual's individual's sensitive information may be disclosed even when data are anonymized using a large k. Following example is conversion of table (a) to (b) and (c) two anonymized tables.

Id	Postcode	Expense (K)
t_1	NW10	*10*
t_2	NW15	*10*
t_3	NW12	*10*
t_4	NW13	*10*
t_5	NW20	*20*
t_6	NW30	*40*
t_7	NW30	*40*
t_8	NW25	*30*

(a)

Id	Postcode	Expense (K)
t_1	*	*10*
t_2	*	*10*
t_3	*	*10*
t_4	*	*10*
t_5	*	*20*
t_6	*	*40*
t_7	*	*40*
t_8	*	*30*

(b)

Id	Postcode	Expense (K)
t_1	NW[10–15]	*10*
t_2	NW[10–15]	*10*
t_3	NW[10–15]	*10*
t_4	NW[10–15]	*10*
t_5	NW[20–30]	*20*
t_6	NW[20–30]	*40*
t_7	NW[20–30]	*40*
t_8	NW[20–30]	*30*

(c)

Figure 1 Data Anonymization – Generalization Algorithms (Li Xiong)

2. l-diversity – It was proposed to prevent value disclosure. Where value disclosure involves the inference of an individual's value in sensitive attribute (SA). This principle requires each anonymized group in T to contain at least *l well represented* SA values.
The simplest interpretation of *well represented* is *distinct* and leads to principle called *distinct l diversity* which requires each anonymized group to contain at least *l* distinct SA values.

3. (a,k) anonymity and p-sensitive-k-anonymity – These principles limit the number of distinct SA values in an anonymized group but still allow an attacker to conclude that individual is likely to have a certain sensitive value, when that value appears much more frequently than others in the group. A principle called *(c,l) – diversity* addresses this limitation.

Achieving k-anonymization with minimum information loss is an NP-hard problem thus many methods employ heuristic search strategies to form k-anonymous groups.

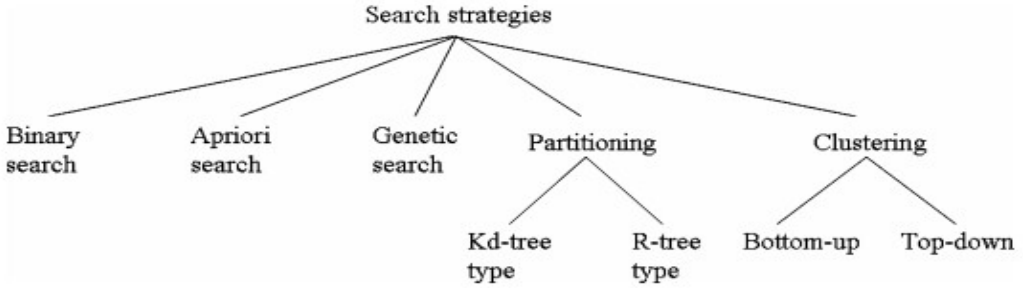


Figure 2 Data Anonymization – Generalization Algorithms (Li Xiong)

Risk Management for HealthCare Data Anonymization -
The factors to consider while sharing Anonymized data are -
1. The likelihood of re-identification for each individual *i*, *li* due to previously released versions.
2. The severity of potentially released sensitive information is if reidentification actually happens.
3. The total expected utility (*u*) of sharing anonymized data.
4. The part of the population *poi* that could be affected by re-identification.

Using above risk factors we can easily define the potential risk in sharing anonymized data as follows -

$$Risk = \bigcup_{i=1}^{m} (s_i, l_i, po_i)$$

$\bigcup$ here is a monotonically increasing function that combines the inputs and transforms them into a standardized range.

Partitioning of datasets for Recoding -
In partitioning we convert a huge dataset to smaller ones for efficient recoding. It is of two types-
1. Single Dimensional – Define non overlapping single dimensional intervals that cover Dxi and map the defined function to each element of Dx.
2. Strict Multi-Dimensional – Define non-overlapping multidimensional intervals that covers Dx1.... Dxd and map the defined function to each element of Dx1.... Dxd.

Recoding of partitioned datasets for Anonymization -
Recoding is changing original value to some range or applying a function over a set of attributes to hide original values. There are two types of Recoding -
1. Global Recoding – Mapping domains of quasi-identifiers to generalized or altered values using a single function.
If Dxi is the domain of attribute Xi in table T. Then
if apply function Fi : Dxi -> D' for each attribute Xi of QIDs.
Global recoding can be single or multidimensional. For example –

Aggregated Table Attributes
Age : 25,26,27,28
Sex : Male,Female
Zip : 53711,53710,53712
Disease: Flu, Hepatitis, Brochitis
Broken Arm , AIDS , Hang Nail

Patient Data

Age	Sex	Zipcode	Disease
25	Male	53711	Flu
25	Female	53712	Hepatitis
26	Male	53711	Brochitis
27	Male	53710	Broken Arm
27	Female	53712	AIDS
28	Male	53711	Hang Nail

Figure 3.1 : Ref - A risk management framework for healthcare data anonymization (Tyron Grandison)

Single Dimensional Partitions
Age : {[25-28]}
Sex: {Male, Female}
Zip : {[53710-53711], 53712}

Multi Dimensional Partitions
Age: [25-26],

Sex: Male,
Zip: 53711}

{*Age*: [25-27],
Sex: Female,
Zip: 53712}

{*Age*: [27-28],
Sex: Male,
Zip: [53710-53711]}

Age	Sex	Zipcode	Disease
[25-26]	Male	53711	Flu
[25-27]	Female	53712	Hepatitis
[25-26]	Male	53711	Brochitis
[27-28]	Male	[53710-53711]	Broken Arm
[25-27]	Female	53712	AIDS
[27-28]	Male	[53710-53711]	Hang Nail

Figure 3.2 : Ref - A risk management framework for healthcare data anonymization (Tyron Grandison)

Age	Sex	Zipcode	Disease
[25-28]	Male	[53710-53711]	Flu
[25-28]	Female	53712	Hepatitis
[25-28]	Male	[53710-53711]	Brochitis
[25-28]	Male	[53710-53711]	Broken Arm
[25-28]	Female	53712	AIDS
[25-28]	Male	[53710-53711]	Hang Nail

Figure 3.3 : Ref - A risk management framework for healthcare data anonymization (Tyron Grandison)

Role Based Access Control (RBAC) - Access control is a means by which the ability is explicitly enabled or restricted in some way.
With Role-Based access control access decisions are based on the roles that individual users have as a part of a system.
It introduces us to the concept of *role* and *permission.* Permissions vary as per the roles. Permissions are assigned to roles.
In terms of healthcare database it is essential to maintain RBAC for patients as it's of utmost concern *who access which information of whom*.
RBAC supports 3 secuirty principles -
1. Least Privilege
2. Separation of duties
3. Data Abstraction
The degree to which data abstraction is supported will be determined by the implementation details of healthcare data.
We can implement RBAC using Oauth, XACML or SAML but Oauth stands steady against the modern technology stack unlike XACML or SAML.

Fast HealthCare Interoperability Resources -
It is a standard for exchanging healthcare information electronically. FHIR must be abided

by to ensure scalability and usability of healthcare data.
HL7 (Health Level 7) has been addressing a few challenges by producing healthcare data exchange and information modelling standards for over 20 years. FHIR is a well documented schema for Healthcare data consistent with the requirements and challenges and the RIM (Reference Information Model), large pictorial representation of clinical data that identifies the cycle that a message or groups of related messages will carry and CDA(Clinical and Administrative Domains) , Document Markup standard.

FHIR consists of several schemas aka. Resources. For all Resources there are elements and properties defined as ID, meta-data , base language and implicit rules.

Problem Formulation

To enable larger scale research, scientists need to share private data collections. For other analysis purposes many organizations require sharing of data associated with some identity. De-identification and Re-identification has primary usage in healthcare data.
Not only the loss of healthcare data harms the integrity of an individual but it may also harm the social persona of him if the data reveals socially awkward information.

It can be used to reduce privacy risks in a wide variety of stuations -
1. Extreme de-identification is used for educational materials that will be made widely public yet must convey enough detail to be useful for medical education purposes.
E.g. Winsconsin Breast Cancer Datasets.
2. Public health uses anonymized databases to track and understand diseases.
3. Clinical trails use it to both protect privacy and subconscious bias by removing other information such as whether the patient received a placebo or an experimental drug.
4. Slight de-identification is used in many clinical reviews, where the reviewers are kept ignorant of the treating physician, hospital, patient, etc. Both to reduce privacy risks and to remove subconscious biases.

Except healthcare industry de-identification is widely adopted by individuals and organizations like e-commerce and big data. With emergence of social network the collected data is de-identified to protect users' privacy. The online shopping websites should adopt this method as well.

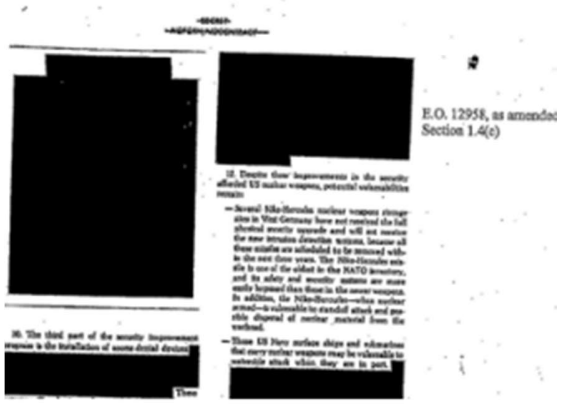
Objectives

The objectives of de-identification and pseudonymization include the following-
1. Exporting anonymized data for secondary use for healthcare data.
2. Building a client side API for access of anonymized data on which the hospital's/ clinic's / Pathology's management system would work.
3. Building an efficient anonymization mechanism that doesn't require batch-anonymization of data before exporting.
4. Ability of re-identification of patients in dire conditions.
5. Continuous public health monitoring and assessment through batch classification of real-time data.
6. Confidential patient-safety reporting (e.g. adverse drug effects)

Methodology

Methods involved in pseudonymization and anonymization of data -

1. Redaction -
It is the process of removing one or more values so that the original information content is no longer obserable by human and computer recipients of the data. Redaction is a type of substitution.
e.g.Historically this technique has been used for legal and governmental work when printed content is physically obscured with a black mark preventing the original content from being read.



Physically redacted USA CIA document.

Figure 4.1 IHE Handbook for Pseudonymization of Healthcare data

2. Fuzzing – It adds apparently random modifications to data while remaining within certain constraints. e.g. a random amount of time can be added to or removed from person's birth date. The goal of fuzzing is to remove as much accuracy as possible while still meeting the intended use.

First_Name	Original_DOB	Fuzzed_DOB	Change
Joe	1997-03-13 13:12	1997-03-14 13:12	Added 1 day
Jane	2005-04-13 11:23	2005-04-10 10:23	Subtracted 3 days and 1 hour
John	1999-06-26 21:24	1999-06-21 19:24	Subtracted 5 days and 2 hours
Pete	2007-10-15 03:13	2007-10-15 06:28	Added 3 hours 15 minutes
Fred	1941-05-16	1941-07-01	Changing month/day to 07/01 preserves year of birth and annual statistics.

Figure 4.2 IHE Handbook for Pseudonymization of Healthcare data

3. Generalization - Generalization is a simpler algorithm than fuzzing but does not preserve statistical characteristics. Several techniques are commonly employed with tradeoffs -

1. reducing precision by truncation of decimal.
2. Generalization of gecodes to a city or landmark.
3. Converting Dates to month number / week number. Etc.

First Name	Original DOB	Fuzzed DOB	Technique Applied
Joe	1997-03-13 13:12	1997-03-14	Removed time
Jane	2005-04-13 11:23	2005-04	Removed day and time
John	1999-06-27 21:24	26	Changed representation to a week of the year number
Pete	2007-10-27 03:13	2000-01-01	Applied a floor (minimum age)
Katie	1923-03-27 14:00	1940-01-01	Applied a ceiling (maximum age)

Figure 4.3 IHE Handbook for Pseudonymization of Healthcare data

4. Longitudinal Consistency Constraints -

It is often essential to preserve date/time relationships, order number relations etc. When the inteded use will examine many related data records preserving these relationships may be important. We refer to this objective as “longitudinal consistency”.

This constraint affects both fuzzing and generalization algorithms.

5. Recoverable Subsitution -

If the original values recoverability is an essential issue then two basic approaches can be used to solve this problem -

1. Escrow – It is widely used in clinical trails. It means holding important information to third party on behalf of transacting parties. e.g. The most common example is replacement of an original 565 patient ID and issuing hospital ID with a clinical subject ID and a clinical trial ID.
2. Encryption of original information.

6. Text Processing and Pass Through -

Text processing aims at de-identification of natural text where as pass through is the data that must be preserved thus will be passed without modification.

References

1. Integrating HealthCare Enterprise De-identification Handbook (*IHE IT Infrastructure Technical Committe)*

2. Overview of Patient Data anonymization (A. Gkoulalas Dvanis and G Loukides *Anonymization of Electronic Medical Records to support Clinical Analysis)*

3. A Risk management framework for healthcare data anonymization (Tyrone Grandison *IBM Services Research ,* Murat Kantarcioglu *University of Texas at Dallas)*

4. Data Anonymization – Generalization Algorithms (*Li Xiong, Slawek Goryczka CS573 Data privacy and Anonymity)*