

# **Novel Solution for Privacy Preservation in E-Healthcare**

## **Data Security**



By

Syed Adeel Shah

A thesis submitted to the faculty of Information Security Department, Military College of Signals, National University of Sciences and Technology, Rawalpindi in partial fulfillment of the requirements for the degree of MS in Information Security

April 2019

## **ABSTRACT**

Privacy preservation is one of the key roles from security perspective in any data security environment. The purpose of this thesis is to provide privacy aspect of security in such a way that it provides strong Patient Anonymity Level, Anonymized Data Searching and successful Correlation of PHR for Medical Research in a single framework. Moreover, a novel solution for data de-identification is introduced (i.e., L-Diversity along with K-Anonymity) for anonymized data searching because previous method of using K-Anonymity (alone) is vulnerable to two type of attacks (homogeneity attack and background knowledge attack). Furthermore, it is experimentally proved in this research that using K-Anonymity alone can risk the disclosure of a huge number of medical records compared to L-Diversity along with K-Anonymity. The percentage risk analysis results are verified as well by using another dataset. Lastly, the experimental setup meets the requirements of HIPAA Privacy Rule as the attributes used in this research are specified by HIPAA as identifying attributes. These identifying attributes are totally suppressed (hidden) as per HIPAA requirement.

## **DECLARATION**

I hereby declare that no portion of work presented in this thesis has been submitted in support of another award or qualification either at this institution or elsewhere.

---

(Syed Adeel Shah)

## **ACKNOWLEDGMENTS**

All praises to Allah Almighty for the strengths and His blessing in completing this thesis.

I would like to convey my gratitude to my supervisor, Dr. Haider Abbas, for his supervision and constant support. His invaluable help of constructive comments and suggestions throughout the experimental and thesis works are major contributions to the success of this research.

I would also like to express my sincerest appreciation to Dr. Rabia Latif and Asst Prof Mian Muhammad Waseem Iqbal for being an important part of my Research Supervisory Committee. Their scholarly guidance, assistance and knowledge have been meaningful for successful completion of my research.

Finally, I am grateful and thankful to Military College of Signals and National University of Sciences and Technology for providing me a chance to help achieve excellence by being associated with the prestigious institutions.

# TABLE OF CONTENTS

<b>Abstract.....</b>	<b>i</b>
<b>Declaration.....</b>	<b>ii</b>
<b>Acknowledgments .....</b>	<b>iii</b>
<b>Table of Contents .....</b>	<b>iv</b>
<b>List of Figures.....</b>	<b>viii</b>
<b>List of Tables .....</b>	<b>ix</b>
<b>Abbreviations .....</b>	<b>x</b>
<b>CHAPTER 1 - Introduction.....</b>	<b>1</b>
1.1 Introduction.....	1
1.2 Motivation and Problem Statement .....	2
1.3 Research Objective .....	3
1.4 Significance for Industry and Military.....	3
1.5 Thesis Contribution.....	3
1.6 Thesis Organization .....	4
1.7 Conclusion .....	4
<b>CHAPTER 2 – Literature Review.....</b>	<b>5</b>
2.1 Introduction.....	5
2.2 Privacy .....	5
2.2.1 Privacy Protection Technologies/Mechanisms .....	5
2.3 Existing Privacy Preservation E-Healthcare Data Security Frameworks .....	6
2.3.1 Yang et. al’s Privacy Preservation Framework.....	6

2.3.2	Agarwal and Johnson’s Privacy Preservation Framework .....	7
2.3.3	Alhaqbani and Fidge’s Privacy Preservation Framework .....	7
2.3.4	Riedl et. al’s Privacy Preservation Framework .....	8
2.3.5	Jian Wang et. al Privacy Preservation Framework .....	8
2.3.6	Pommering et. al Privacy Preservation Framework .....	9
2.3.7	Aamot et. al Privacy Preservation Framework .....	9
2.3.8	Adeela Waqar et al. Privacy Preservation Framework .....	10
2.3.9	Win KT et al. Privacy Preservation Framework .....	10
2.3.10	Narayan et al. Privacy Preservation Framework .....	10
2.4	Considered Parameters for This Research .....	11
2.4.1	Patient Anonymity Level .....	11
2.4.2	Correlating PHR for Medical Research .....	11
2.4.3	Anonymized Data Searching .....	12
2.5	Flaws/Vulnerabilities in existing frameworks .....	13
2.6	K-Anonymization, L-diversity analysis for the use in the E-Healthcare framework .....	15
2.6.1	K-Anonymization .....	15
2.6.2	Why not using k-Anonymity .....	16
2.6.3	L-diversity.....	16
2.7	Terms for understanding the privacy preservation about anonymized data searching .....	16
2.7.1	Microdata .....	16
2.8	Conclusion .....	17

<b>CHAPTER 3 –Proposed Framework For Privacy Preservation In E-Healthcare Data Security.</b> .....	<b>18</b>
3.1 Introduction.....	18
3.2 Privacy Preservation Framework & Considered Parameters.....	20
3.2.1 Patient Anonymity Level .....	21
3.2.2 Correlation of PHR for Medical Research.....	21
3.2.3 Anonymized Data Searching .....	21
3.3 Semantic Validation.....	22
3.3.1 Attacks On k-Anonymity .....	24
3.4 L-Diversity Postulate .....	26
3.5 HIPAA Privacy Rule & Classification of Attributes .....	27
3.6 Conclusion .....	28
<b>CHAPTER 4 - Implementation &amp; Results.....</b>	<b>30</b>
4.1 Introduction.....	30
4.2 Experimental Setup.....	30
4.2.1 ARX Anonymization Tool.....	30
4.3 Case Scenarios .....	31
4.3.1 2-Anonymity .....	31
4.4 Results.....	34
4.4.1 K-Anonymization Plots .....	34
4.4.2 K-Anonymity with L-Diversity Plots .....	36
4.5 For Verification of Results.....	39
4.5.1 2-Anonymity .....	39

4.6	Conclusion .....	42
<b>CHAPTER 5 – Conclusion &amp; Future Directions.....</b>		<b>44</b>
5.1	Conclusion .....	44
5.2	Future Directions .....	44
5.3	Summary .....	45
<b>References .....</b>		<b>46</b>



## LIST OF FIGURES

Figure 2.1: Pseudonymization based on reported results flow direction .....	9
Figure 2.2: AES-128, 192 or 256 bit.....	12
Figure 2.3: Federated Identity Management Architecture .....	13
Figure 3.1: Proposed Framework.....	18
Figure 3.2: K-Anonymity Vulnerabilities.....	22
Figure 4.1: 2-Anonymity .....	32
Figure 4.2: 2-Anonymity Input and Output EMRs.....	32
Figure 4.3: Risk Analysis before 2-Anonymity .....	33
Figure 4.4: Risk Analysis after 2-Anonymity.....	33
Figure 4.5: K-Anonymization values w.r.t. Percentage of Records at Risk before Anonymization .....	35
Figure 4.6: K-Anonymization values w.r.t. Percentage of Records at Risk after Anonymization .....	36
Figure 4.7: K-Anonymization & L-Diversity values w.r.t. Percentage of Records at Risk before Anonymization .....	37
Figure 4.8: K-Anonymization & L-Diversity values w.r.t. Percentage of Records at Risk after Anonymization .....	38
Figure 4.9: 2-Anonymity .....	39
Figure 4.10: Risk Analysis before applying any Anonymization Technique.....	40
Figure 4.11: Risk Analysis after 2-Anonymity.....	40

## LIST OF TABLES

Table 2.1: Strengths & Weaknesses of Existing Frameworks .....	13
Table 3.1: Quasi-identifiers .....	19
Table 3.2: Voter Registration Records.....	20
Table 3.3: Anonymized Medical Dataset.....	20
Table 3.4: Medical Records Table .....	23
Table 3.5: 4-Anonymous Table .....	24
Table 3.6: 3-Diversity .....	26
Table 4.1: Percentage of Risk Analysis on different Anonymization Values .....	34
Table 4.2: K-Anonymization Values w.r.t. Percentage of Records at Risk before Anonymization .....	34
Table 4.3: K-Anonymization Values w.r.t. Percentage of Records at Risk after Anonymization .....	35
Table 4.4: K-Anonymization and L-Diversity Values w.r.t. Percentage of Records at Risk before Anonymization.....	37
Table 4.5: K-Anonymization and L-Diversity Values w.r.t. Percentage of Records at Risk after Anonymization.....	38
Table 4.6: Percentage of Risk Analysis on different Anonymization Values .....	41
Table 4.7: Percentage of Input Risk Analysis of both Datasets.....	41
Table 4.8: Percentage of Output Risk Analysis of both Datasets .....	42

## ABBREVIATIONS

EHR	Electronic Health Record
IT	Information Technology
PHI	Protected Health Information
EPR	Electronic Patient Record
PR	Patient Record
BMA	British Medical Association BMA
FIP	Fair Information Practices
PHR	Personal Health Record
MAC	Mandatory access control
RBAC	Role-based access control
HDB	Hippocratic Database
PPDM	Privacy Preserving Data Mining
SII	Sovereign information integration
EMRs	Electronic Medical Records
FIM	Federated Identity Management
EHS	Electronic healthcare system
HIPAA	Health Insurance Portability and Accountability Act
CP-ABE	Ciphertext-Policy Attribute-Based Encryption
TA	Trusted Authority
PEKS	Public Key Encryption with Keyword Search

SSL	Secure Sockets Layer
AES	Advanced Encryption Standard
PSNs	Pseudonyms
IDP	Identity Provider
SP	Service Providers
SSO	Single Sign-on

## **INTRODUCTION**

### **1.1 Introduction**

Electronic healthcare (E-Healthcare) technology is becoming one of the most populous and promising technologies in today's world. Old methods of maintaining medical records usually on paper is getting obsolete because in today's modern world, it was not sufficient to fulfil the needs of contemporary treatment which makes numerous medical organizations and academia shifting from paper based records to electronic healthcare systems which can be managed efficiently and easily [11].

Electronic Health Record (EHR) is the file structure for communication of health information. IT groundwork that enables EHRs sharing precisely is called EHR system. EHR system illustrates the usability of information and communication technologies across the whole new level which directly affect Protected Health Information (PHI). Quality of diagnosis can be improved tremendously because of the huge potential of E-Healthcare system. Moreover, it has the tendency in the reduction of medical costs and provides the assistance to address the reliable and on-demand healthcare tests which can be posed by the aging society [12].

During the healthcare process, patient's data is gathered known as patient record (PR). It could be paper based or electronically collected. If the patient's data is stored electronically, that is known as electronic patient record (EPR). EPRs provides facility for patients records to be transferred from paper based to electronic records which can then be kept digitally. E-Healthcare systems possess the tendency to transform healthcare facilities to a more practical and user oriented framework of healthcare which will help in not only improving cost, quality but also accessibility of healthcare services. Using EPRs in comparison to paper based reduces mistakes in healthcare by allowing patients to track the progression of treatment. Moreover, it will provide better and secure links among patients and e-healthcare providers [31].

"Patients can have additional and improved access to their data" is the most beneficial property of EPR in comparison with paper-based records [13]. Through using EPR, patients have

additional and flexible check over their health data. Moreover, EPRs have tendency to ease patients in tracking their own illness progress [14]. The prime aim of EPR system is to offer an environment for patients where they can securely and safely exchange and share their records.

According to EU Directive [15], privacy makes its place among the elementary and primary human rights. While on other hand, privacy defined by [19] in context of E-Healthcare that it is the desire and right of patient to control and manage the sharing and disclosure of his personal health records. Disclosure is the unveiling or releasing of patient's identifiable attributes to others referred from British Medical Association (BMA) Ethics [23]. Fair Information Practice (FIP) [24] define privacy as an approach to prevent somebody from getting a person's information for one motive and after the completion of that motive, it is utilized for some other intention without the individual's assent. Building up legitimate protection and security laws to characterize patient's rights to uncover information by patient's assent is important. As indicated by [25], patients need to be kept aware of their healthcare records. They must be well informed that how is their data saved and whose authorized to access their personal health data and for what motive. Specialized and authoritative procedures must be fulfilled by data manager to secure user information to accidental or unlawful loss, alternation, unauthorized entrance over network or information disclosure, and against all further illegal exercises.

This chapter is organized as accordingly: Section 1.2 highlights the research significance. In section 1.3, the motivation for doing this research along with the problem statement has been described. In section 1.4, objectives of this research are briefly discussed. Section 1.5 shows the contributions made by this research in the field of Information Security. Finally, Section 1.6 gives an overview of other chapters included in this research.

## **1.2 Motivation and Problem Statement**

Currently, not a single privacy preservation E-Healthcare data security framework provides strong patient anonymity level including both patient identity and data, anonymized data searching & successful correlation of Personal Health Record (PHR) for medical research. Moreover, k-anonymity is the only technique used for anonymized data searching currently and k-anonymity has some existing flaws discussed in the later chapter which can be exploited.

Therefore, there is a need of a privacy preservation E-Healthcare data security framework which has capability to provide all these three components i.e., strong patient anonymity level, anonymized data searching & successful correlation of PHR for medical research in a single framework. Furthermore, it should provide immunity against the vulnerabilities posed by k-anonymity.

### **1.3 Research Objective**

Objectives for this research are as follows:

- Perform analysis, comparison & discussion on the existing privacy preservation E-Healthcare frameworks.
- Point out flaws in the existing privacy preservation E-Healthcare frameworks and its solutions.
- Propose an enhanced solution for privacy preservation E-Healthcare framework.
- Validation of proposed framework.

### **1.4 Significance for Industry and Military**

This research will focus the existing flaws in current E-Healthcare frameworks and it will provide an innovative solution for privacy preservation in E-Healthcare framework where it will mitigate the threats of existing attacks on current E-Healthcare frameworks. Moreover, this framework will be using a new technique for data de-identification which proves better than the former method of data de-identification.

This research carries immense importance in context of E-Healthcare data security of Pakistan. Currently, Pakistan hasn't faced any major healthcare data security breach but it is wiser to tackle against these threats beforehand as it can compromise a major loss of capital as well as patient's trust as we have seen such attacks in other parts of globe which has suffered from these threats. This research exclusively targets the security of healthcare industry by providing a privacy preservation framework which will enhance its overall security.

### **1.5 Thesis Contribution**

This research provides multiple contributions as follows:

- Provides an overview of existing privacy preservation E-Healthcare Data Security.
- Performed analysis, comparison and discussion on multiple privacy preservation E-Healthcare frameworks.
- Comparative Analysis of K-Anonymity and L-Diversity through ARX Anonymization tool with risk analysis.

## 1.6 Thesis Organization

This thesis is divided into five chapters as follows:

- **Chapter 1** covers introduction, significance of this research, motivation and problem statement, objectives and thesis contribution.
- **Chapter 2** is about literature review done during this research. Former privacy preservation frameworks are presented and critically analyzed with respect to considered parameters for this research. Flaws in those frameworks are identified and data de-identification techniques are discussed.
- **Chapter 3** provides proposed framework for privacy preservation in E-Healthcare Data Security.
- **Chapter 4** validates the proposed framework with the help ARX Anonymization Tool via developing various case scenarios and deriving results through them.
- **Chapter 5** gives the conclusion and future direction.

## 1.7 Conclusion

The objective and motivation to conduct the research on privacy preservation on E-Healthcare data security has been described in this chapter. The research objectives of this research are also mentioned. Its importance for industry and military has also been highlighted. At the end it describes the overall structural organization of the thesis.



## **LITERATURE REVIEW**

### **2.1 Introduction**

In this chapter, we have highlighted the literature review done for this research. Existing privacy preservation frameworks are discussed and analyzed in detail. Vulnerabilities in the existing frameworks are also pointed out with respect to the parameters considered for this research namely, patient anonymity level, correlation of PHR for medical research and anonymized data searching. Moreover, anonymization techniques are discussed and critically analyzed namely, K-Anonymity and L-Diversity. Finally, some terms in reference to privacy preservation are also explained for the ease in understanding.

### **2.2 Privacy**

Privacy according to Alan Westin [16] is defined as “the right of entities, groups or organizations to regulate for themselves, when, how and to what extent information about themselves is disclosed to public”. According to this definition, person (entities) as well as legal organizations (groups or institutions) have the right to privacy. In this research we will be considering privacy in context of informational privacy, which controls whether and how personal information can be collected, stored, administered or selectively communicated.

#### **2.2.1 Privacy Protection Technologies/Mechanisms**

Privacy protection technologies refer to a variety of technologies that protect personal privacy by minimizing or discarding the collection of recognizable information [17].

The privacy protection technologies cover a variety of features, such as:

- Safeguarding the user personal identity with anonymization, pseudonymization, unobservability of users and unlinkability. The legal postulate of requirement of collecting the data and managing it requires that personal information must only be gathered or used for identification purposes only when truly necessary. If personal data has to be collected, it should be used anonymous or pseudonymous as soon as the motive for which the information was collected permits that.

- Using access control mechanisms confidentiality, integrity, and availability of personal information can be protected. The privacy requirements of necessity of data processing of personal data of users and data subjects, which requires that access to personal data is necessary for performing current task, and purpose binding, which requires that the purpose of the access should be, through access control mechanisms and an appropriate security policy can be technically enforced.

## **2.3 Existing Privacy Preservation E-Healthcare Data Security Frameworks**

In this section, some existing privacy preservation frameworks are described which will be further elaborated in section 2.5 discussing their strengths and weaknesses.

### **2.3.1 Yang et. al's Privacy Preservation Framework**

In [1], author presented three well known and advanced techniques in order to ensure privacy in E-Healthcare data security environment namely privacy by policy, privacy by cryptography and privacy by statistics. Author made a hybrid model where statistical analysis and cryptography were incorporated which strengthens several models of health information to an ample degree of privacy strength. Sharing and exchanging of health information between various healthcare providers is also discussed here. Dataset-level privacy security is the main focus for their solutions [33]. Furthermore, author discussed here the issue of sharing of health records while keeping privacy of health records preserved that how are these records shared in the cloud computing [34]. Mandatory access control (MAC) and Role-based access control (RBAC) are the only conventional access control models which are proposed by researchers where only policy for privacy preservation is concerned [35]. However, these two access control models barely fulfill the privacy preservation requirements as of the absence of basic elements that are needed for preserving privacy. However, privacy by statistical analysis's proposed technologies provides strong guarantee where privacy is concerned but merely towards limited attack model, and this is not acceptable as it cannot limit the attacker who might possess rich background knowledge or has access to considerate amount public information. Considering cryptography here, it provides a theoretical guarantee as it has tendency to fulfill strong privacy preservation requirements, however, where information utility is concerned, cryptography offers limited access pattern. The author [1], illustrated these three privacy preservation techniques here namely privacy by statistics,

privacy by cryptography and privacy by policy. Author then proposed his hybrid approach for preserving privacy as it offers all the attributes from former techniques.

### **2.3.2 Agarwal and Johnson's Privacy Preservation Framework**

Agarwal's [2] proposal for preserving privacy in E-Healthcare environment was constituted of a Hippocratic Database (HDB) technique. This technique comprises of five elementary components namely; data mining algorithm, efficient auditing (compliance auditing), active enforcement, sovereign information integration (information sharing) and an ideal k-anonymization.

Firstly, for the screening process of SQL queries, active enforcement component is used which is an efficient approach to offer patients to their preferences. Secondly, to track the security breaches, efficient auditing component of HDB is used. Moreover, through efficient auditing, it can check whether policies are in fully compliant by auditing past database access. Third comes the data mining algorithm which is the program for monitoring health conditions from home. Patient's home computer is part of this component where this health data is recorded and two copies of this record are made. One copy is sent to a randomizer known as Privacy Preserving Data Mining (PPDM) while the second copy is sent to database of patient's hospital. Here randomizer plays a vital job as it satisfies patient's preference policy by randomizing the data. Forth component which is quite important as it is used to de-identify the personal health data by using k-anonymization. Generalization and suppression are the two techniques involved in k-anonymity. Direct identifiers are fully suppressed whereas other attributes are generalized that could be used for research purposes. Finally, Sovereign information integration (SII) component uses cryptographic protocols in order to secure data. SII is quite handy as it will disclose only results of the applied query. Moreover, it is scalable as well.

### **2.3.3 Alhaqbani and Fidge's Privacy Preservation Framework**

Alhaqbani et. al's in [3], introduced a framework which has the tendency to provide patients full control over their health information. This could be achieved by joining Electronic Medical Records (EMRs). For the fulfilment of this job, indirect pseudonym identifiers are used. The EHRs are then linked to their respective distinct EMRs. A unique identifier is assigned to every individual EMR which will later be used for making any corrections of patient's unique ID.

This unique identifier is already present in the database of healthcare provider. Author illustrated in this paper two techniques of identity linkage and present record. Moreover, they demonstrated that these techniques are insufficient to meet the privacy and accuracy requirements for EHR systems. They also pose a solution for this problem which was the usage of indirect pseudonym identifiers along with records of patients that are associated to an EHR identity-management architecture. Using such method satisfies the privacy requirements and it yields precise results as per existing Federated Identity Management (FIM) architecture.

#### **2.3.4 Riedl et. al's Privacy Preservation Framework**

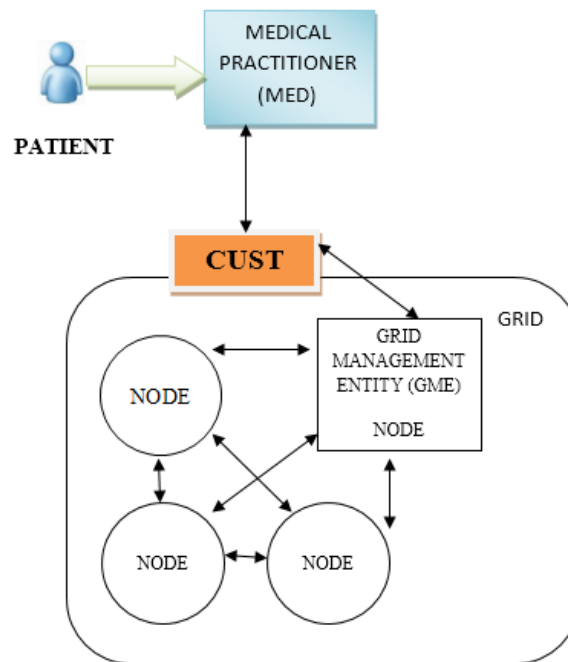
Riedl et. al [5], to prevent two pseudonyms which can originate from a single PHI, the author intended to assign new pseudonym to a PHI for each session. User's information, if it does not disclose any information about its direct identifiers, then it can be concluded that it satisfies the user anonymity requirement [21]. The identity of patient must remain secure, more elaboratively, it can be said that the patient's ID must remain anonymous in electronic healthcare system (EHS). One suggested approach to achieve this anonymization constraint is through pseudo anonymity. Via pseudo anonymity, a trusted third party accesses patient's records and assign a value to that identifier, which works like a hash function as patient's original identity cannot be traced back. Shamir's threshold scheme [20] was the basis of Riedl et al.'s proposition for sharing pseudonyms. Moreover, lost or destroyed keys could also be recovered through this mechanism.

#### **2.3.5 Jian Wang et. al Privacy Preservation Framework**

Preserving privacy in cloud computing, Jian Wang et al. [6] put forward his anonymity-based approach. In this work, a small sample of data was used to illustrate the anonymization which was based on service provider's background knowledge. However, this work can only be applicable on a very limited services and also it was not sufficient in this type of anonymization in order to automate it. This approach used is quite simple and flexible as it does not gets freed from key management system which is a technique used in traditional cryptography for preserving privacy of user. In their work, they also put forward their idea of data anonymization which can be used along with data partitioning.

### 2.3.6 Pommering et. al Privacy Preservation Framework

Two-tiered pseudonymization was proposed by Pommering et. al in [4]. In [17], it is quite well explained. Figure 2.1 provides a diagrammatic elaboration. Management of identifiers also known as pseudonyms is done by custodian (CUST). CUST act as an intermediary body between medical practitioner and grid management entity. Pseudonymization of healthcare information is done by CUST after which data is transferred to grid management entity and grid nodes from medical practitioner. Whereas the de-pseudonymization process is done when the data is sent back to the medical practitioner.



**Figure 2.1: Pseudonymization and de-pseudonymization process on the basis of data stream direction**

### 2.3.7 Aamot et. al Privacy Preservation Framework

Julia Bickford and Jeff Nisker in [18] explained Aamot et. al's [7] work concisely. It highlights the complications inflicted with anonymization process. It also discusses the enhancement of pseudonymization process through the support of research ethics board. The enhancement in pseudonymization process, will allow researcher's communication in genomic

research which will directly benefit patients as the results of the research could be used by patients for their own clinical motives.

### **2.3.8 Adeela Waqar et al. Privacy Preservation Framework**

Probability of metadata manipulation is the main focus in Adeela Waqar et al.'s [8] work. The attacker can easily disrupt privacy of a user if he/she manages to have access to the metadata. Author proposed a framework to counter against this problem and have privacy of data preserved. Firstly, metadata is separated to be put in the database of cloud. Then the process of data classification took place in which metadata is sorted with respect to its attributes according to the sensitivity of data. Table splitting process took place after data classification is done successfully. Division of database tables both vertically and horizontally are the main approaches in the process of table splitting. Along with table splitting process, normalization of database is also guaranteed. Then a process known as ephemeral referential consonance took place where metadata is reconstructed as required for the cloud. These processes discussed above provides a guarantee that data from cloud database cannot be compromised both before and after the splitting process. Potential attacks are prevented with this proposed framework. These attacks on metadata were demonstrated considering the above mentioned processes which were saved in Eucalyptus database records.

### **2.3.9 Win KT et al. Privacy Preservation Framework**

Authentication of user in a secure way is the main focus addressed by Win KT et al.'s work [9]. Author accomplishes this user authentication with the help of a trusted authority also known as certification authority. In this approach, user has full control over his health records and only he can access to the records through an authentic credential that is provided by the trusted authority. User has now tendency to perform cryptographic operations such as signing a document or decrypting the information. Interference or interpretation's probability is very low as strong cryptographic operation namely encryption of 128-bit is used which is very hard to crack except for quantum computing. HIPAA's regulations and standards are also obeyed in this work.

### **2.3.10 Narayan et al. Privacy Preservation Framework**

To counter against cloud provider, as there is a possibility that they can breach the privacy and integrity of EHR records in the cloud database, Ciphertext-Policy Attribute-Based Encryption

(CP-ABE) approach was introduced by Narayan et al. [10]. EHR data and metadata is encrypted via asymmetric encryption of public and private keys through the use of attribute-based encryption (ABE) scheme. Key management task is performed by a Trusted Authority (TA). All EHRs that are encrypted can be accessed by TA. For private search of EHRs and security, Public Key Encryption with Keyword Search (PEKS) and CP-ABE scheme are used in combination. Encryption of data is executed via symmetric key cryptography while the access to symmetric keys for authorized users are given through ABE scheme. For protection against attacker to eavesdrop to gain any information about the key, SSL protocol is used which establishes a secure link in order to share the private key safely.

## **2.4 Considered Parameters for This Research**

Three parameters of privacy preservation are considered for this research (apart from enhancement in anonymized data searching) i.e. patient anonymity level, correlating PHR for medical research and anonymized data searching. In further subsections these three parameters are described and how they are achieved is elaborated.

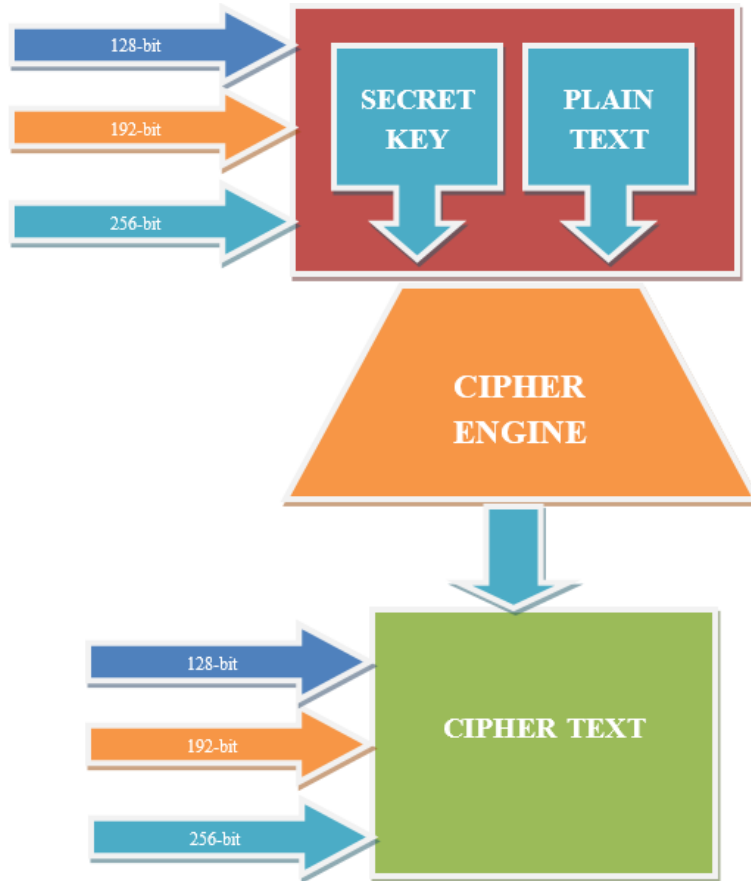
### **2.4.1 Patient Anonymity Level**

This parameter of patient anonymity is of two levels which are low level (with weak or no patient anonymity level) and high level (with strong patient anonymity level). This high level or strong patient anonymity level is achieved through a cryptographic technique of encryption. For much better level of security, it is recommended to use AES-128, 192 or 256 bit as shown in Figure 2.2 [21]. Patient identity (ID) and patient data both can be encrypted and used to enhance the patient anonymity level.

### **2.4.2 Correlating PHR for Medical Research**

Requirement of this parameter can be fulfilled through using pseudonyms (PSNs). Federated Identity Management (FIM) is the body through which PSNs are created. Some basic concepts on FIM are illustrated in Figure 2.3 [18]. Identity provider (IdP) plays its role as FIM's authoritative part. IdP plays an important part in user validation and assigning a trusted identity to user. On that assigned trusted identity, user can interact with its reliable business associates. Services are offered by business associates and the component which carries out this task is known as service providers (SP). SP tasks are reduced considerably as most of the tasks are segregated such as access

management or authentication process of user. Moreover, SP gains sufficient trustworthy information regarding user that now user registration with the SP is also not required.



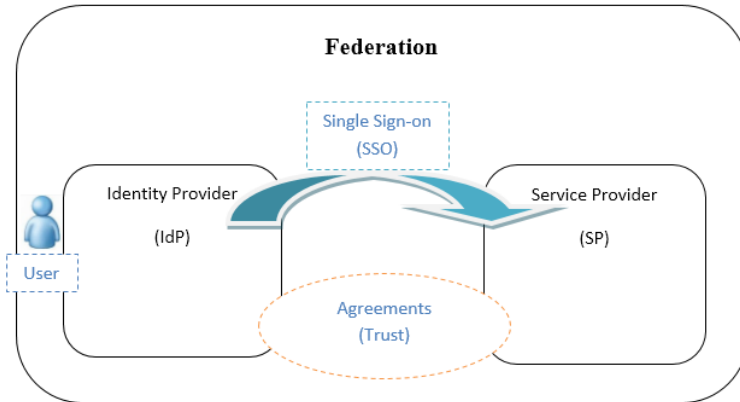
**Figure 2.2: AES-128, 192 or 256 bit**

### 2.4.3 Anonymized Data Searching

With the use of k-anonymization which is an NP-hard crypto problem, successful anonymized data searching parameter can be fulfilled which is data de-identification process. However, there are some existing flaws in using k-anonymity in the E-Healthcare environment which are discussed in detail in section 3.3. Due to these vulnerabilities present in the k-anonymization, we turn towards L-diversity.

This is the most important component for this research as this component of privacy preservation framework is elaborated in detail. Its vulnerabilities are identified, and an enhanced version of data de-identification is used in it.





**Figure 2.3: FIM's Architecture**

## 2.5 Flaws/Vulnerabilities in existing frameworks

**Table 2.1 Strengths & Weaknesses of Existing Frameworks [46]**

<b>Research Article</b>	<b>Patient Anonymity Level</b>	<b>Correlation of PHR for Medical Research</b>	<b>Anonymized Data Searching</b>
Yang et. al [1]	YES	NO	YES
Agarwal et. al [2]	YES	NO	YES
Alhaqbani et.al [3]	NO	YES	NO
Riedl et. al [4]	NO	YES	NO
Jian Wang et. al [5]	NO	NO	NO
Pommering et. al [6]	NO	NO	NO
Aamot et. al [7]	NO	NO	NO
Adeela Waqar et al. [8]	YES	NO	NO
Win KT et al. [9]	YES	NO	NO
Narayan et al. [10]	YES	NO	NO

Author's proposal [1] of hybrid approach present a strong privacy preservation framework as encryption was used which makes patient anonymity level quite strong. Moreover, with the use of k-anonymity, which is considered to be a strong data de-identifying technique, anonymized data searching parameter requirement is also fulfilled. However, correlating PHRs for medical research requirement cannot be met because of not using PSNs.

Based on HDB approach, a strong privacy preservation was proposed by Agarwal and Johnson [2]. In their framework, they also use encryption and k-anonymity which meets the requirement of anonymized data searching and strong patient anonymity level however like [1], it also didn't make use of PSNs through which it was also not able to fulfill the correlation of PHR for medical research parameter.

Unlike [1] and [2], Alhaqbani and Fidge in [3], the parameter of correlation of PHR for medical research property was fulfilled with the use of PSNs. However, the other two parameters of anonymized data searching, and strong patient anonymity level wasn't fulfilled because of the absence of encryption and any de-identification technique.

Usage of PSNs are part of Riedl et. al's [5] work. However, for every session of PHI, a new PSN was assigned makes it quite impractical to correlate PHR for medical research. Moreover, patient anonymity level was also weak and there was no anonymized data searching because of no use of any data de-identification approach.

In the work of Wang et. al [6], with respect to our considered parameters, it failed to provide strong patient anonymity level because of the absence of encryption. It was also not able to provide successful correlation of PHR for medical research and also anonymized data searching parameter requirement was also not met.

Based on two-tiered pseudonymization, Pommering et. al's [4] approach, with the use of PSNs, their framework met the requirement of correlation of PHR for medical research parameter. But like [3], it was also not able to provide strong patient anonymity level and anonymized data searching parameters.

Pseudonymization problems were identified by Aamot et. al's [7] in their work. However, as far as our considered parameters are concerned, it was not able to meet the requirements of all

the parameters such as successful correlation of PHR for medical research, anonymized data searching and strong patient anonymity level.

The framework proposed by Adeela Waqar et al. [8] works quite efficiently. With the use of metadata reconstruction in a dynamic way, cloud user's data privacy is ensured. Encryption technique is part of this framework which gives patient a strong patient anonymity level. However, other two parameters such as anonymized data searching and correlation of PHR for medical research were not fulfilled.

The proposed framework by Win KT et al.'s [9] work provides full control to patients over their EMRs. Also, it offers special services for mentally incapable and minors. Their framework was quite efficient as it provides strong patient anonymity level because of the usage of AES encryption of 128-bit. But, the requirement of other two parameters such as anonymized data searching and correlation of PHR for medical research were not met.

In order to secure EHRs, numerous solutions were offered in the proposed work of Narayan et al. [10]. However, their framework could not be used for large scale access. Their framework also offers strong patient anonymity level because of using encryption. However, the other two parameters requirement was not fulfilled as in their framework they neither used PSNs nor any data de-identification technique.

## **2.6 K-Anonymization, L-diversity analysis for the use in the E-Healthcare framework**

In this section, these two data de-identification techniques (k-Anonymization and L-diversity) are described and are critically analyzed.

### **2.6.1 K-Anonymization**

One of the powerful techniques used in data de-identification is K-Anonymity. K-anonymized dataset has a distinct characteristic that each record is indistinguishable from at least k-1 other records. According to Samarati and Sweeney, sophisticated methods to de-identify microdata are vulnerable to attacks that collect the information with other public sources available data to re-identify the specific entity. In order to avoid these types of linking attacks while keeping the integrity of public information preserved, the notion of k-anonymity [26] was put forward by

Samarati and Sweeney. The value of  $k$  and the implicit privacy is directly proportional as no specific entity can be recognized with likelihood exceeding  $1/k$  through linking attacks alone.

### **2.6.2 Why not using k-Anonymity**

K-anonymity undoubtedly is a strong approach for data de-identification but still it has some severe privacy problems. Two very practical and doable attacks could be launched against k-anonymity which are: firstly, background knowledge (background knowledge attack) and secondly, sensitive attributes with little diversity (homogeneity attack) [22].

So, it can be concluded for k-anonymity that it can protect against identity disclosure but not against attribute disclosure

### **2.6.3 L-diversity**

L-diversity was formally presented by Machanavajjhala et al. [27] which was based on the concept of Bayes-optimal privacy in order to counter the against the possible attacks on k-anonymity. Each equivalence class has not less than  $l$  well-represented values for each sensitive attribute. An amazing benefit of L-Diversity is that it will still fulfil privacy requirement even if the data publisher does not need to be aware of what extent of information the attacker possess. Values of sensitive attributes should be well-represented in each group is the key motive behind L-Diversity.

## **2.7 Terms for understanding the privacy preservation about anonymized data searching**

Privacy is a major issue when microdata is going to be released. The release of microdata incurs apparent privacy concerns. Basic de-identification has been revealed to be inadequate, as privacy can be compromised if quasi-identifier characteristics in a de-identified database are connected/related to publicly accessible/available data. To counter against such attack, generalization and suppression-based approaches (like k-anonymity) have been put forwarded to weaken the association between the quasi-identifiers of a record and its sensitive attributes in a microdata record. It is important to know about certain terms in context of privacy preservation.

### **2.7.1 Microdata**

Microdata is released data in table form with three types of attributes. i.e., identifiers, quasi-identifiers and sensitive attributes.

The attribute identifier (ID) is very crucial attribute whose value if known can specifically identify that entity. There can be multiple identifiers e.g., social security number, mobile or telephone number, CNIC number etc. These attributes are not released at all whether it is in microdata table or in any other publicly available/accessible databases.

The attribute quasi-identifier can be linked with tuples which are records released in the microdata table. These attributes appear in the microdata table however they do not appear in any other publicly available/accessible databases. Zip code, age, DOB etc. are the examples of quasi-identifiers.

Sensitive attributes as the name suggests is of immense importance to its respective entity. These attributes do not appear in publicly available databases however they are included in the microdata table. The main objective of privacy protection here is to prevent attackers from knowing the values of sensitive attributes which can then be used in associating with respective patient record.

## **2.8 Conclusion**

In this chapter we have explained some existing privacy preservation frameworks and pointed out some flaws in them. We have also elaborated the considered parameters for this research namely, patient anonymity level, correlation of PHR for medical research and anonymized data searching. Furthermore, we have also given a brief introduction on the anonymization techniques we will be using in this research namely, K-Anonymity and L-Diversity.

## PROPOSED FRAMEWORK FOR PRIVACY PRESERVATION IN E-HEALTHCARE DATA SECURITY

### 3.1 Introduction

This is the core chapter of this research work as in this chapter two framework will be combined in such a way that it will takes the benefitting characteristics of former frameworks. Considering all three required parameters for our research as shown in figure 3.1 i.e. patient anonymity level, anonymized data searching and correlating PHR for medical research and integrating them all in a single framework. Moreover, a new parameter L-diversity will be introduced along with K-Anonymity in this framework instead of former techniques for data de-identification.

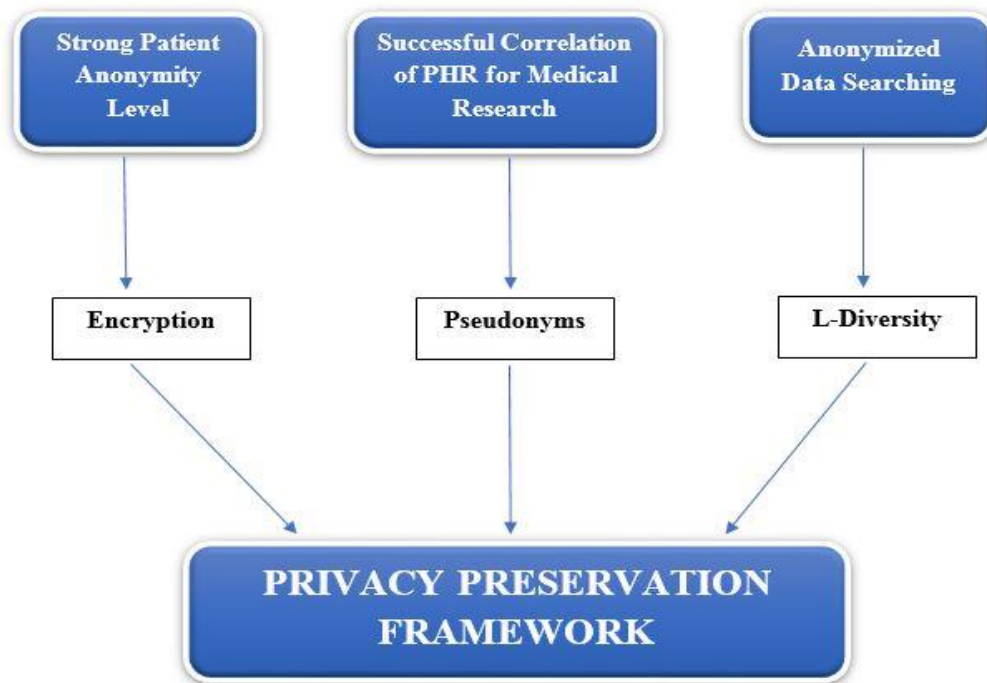


Figure 3.1 Proposed Framework

Microdata publishing is becoming very common now among many organizations. These microdata tables can contain very important and useful data for both attackers and researchers such as customer data, census, voter registration and medical data. This microdata information is a valuable source of information as from this study, different trend analysis and medical researches are being conducted. However, if this microdata can distinctively identify an individual, it can compromise its sensitive attributes which is unacceptable.

To prevent this sort of mishap in which microdata can uniquely identify an entity, social security numbers (SSN) and names are removed from microdata tables. This is known as first level sanitization which does not ensure privacy preservation of individual in the microdata tables. According to a recent study conducted in Carnegie Mellon University [31], 87% people of United States can be distinctively identified only through the usage of simple attributes such as age, gender and a 5-digit zip code, shown in table 3.1 These attributes are quasi-identifier. As a matter of fact, these three attributes were linked to voter registration records shown in table 3.2 of Massachusetts which contains name, zip code, gender and date of birth (DoB). Those three attributes were also linked to an anonymized medical data as shown in table 3.3 of GIC1 which contains zip code, gender, DoB and diagnosis. This type of attack is known as “linking attack” and through this attack, it was managed to distinctively identify Governor of Massachusetts medical record [32].

**Table 3.1: Quasi-identifiers**

	<b>Age</b>	<b>Gender</b>	<b>Zip Code</b>
<b>1</b>	*	*	*
<b>2</b>	*	*	*
<b>...</b>	*	*	*
<b>n</b>	*	*	*

**Table 3.2: Voter Registration Records**

	<b>Name</b>	<b>Gender</b>	<b>Zip Code</b>	<b>Date of Birth</b>
<b>1</b>	*	*	*	*
<b>2</b>	*	*	*	*
...	*	*	*	*
<b>N</b>	*	*	*	*

**3.2 Privacy Preservation Framework and Considered Parameters**

Three privacy preservation parameters are considered for the framework in this research namely patient anonymity level, anonymized data searching and correlating PHR for medical research. In addition, a modification in the data de-identification technique by introducing L-Diversity in the framework as well.

**Table 3.3: Anonymized Medical Dataset**

	<b>Zip Code</b>	<b>Gender</b>	<b>Date of Birth</b>	<b>Diagnosis</b>
<b>1</b>	*	*	*	*
<b>2</b>	*	*	*	*
...	*	*	*	*
<b>N</b>	*	*	*	*



### **3.2.1 Patient Anonymity Level**

For ensuring a strong a patient anonymity level of both data and identity of patient, a cryptographic technique is used i.e., encryption. A much better option in encryption will be using Advanced Encryption Standard (AES) which has a block size of 128 bits, and key lengths of 128, 192 and 256 bits.

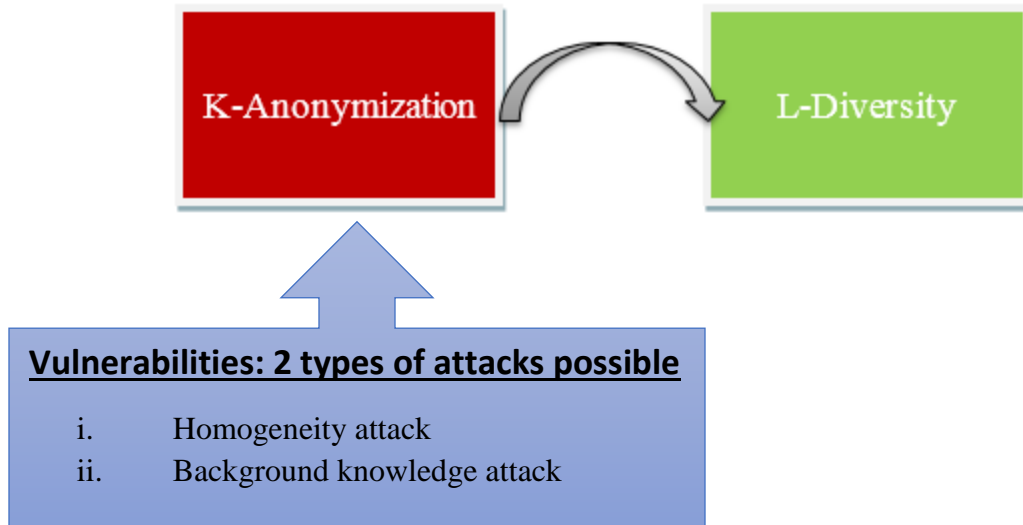
### **3.2.2 Correlation of PHR for Medical Research**

For the parameter of correlation of PHR for medical research, this can be achieved with the use of pseudonyms (PSNs) which are formed by Federated Identity Management (FIM). The Figure 3.2 below shows some fundamental thoughts in federated identity management (FIM) [18]. The authoritative part of FIM is the identity provider (IDP) which is accountable for validating the user and declaring a trusted identity for the user to its reliable business associates. Business associates who would offer services but would not provide identities are recognized as service providers (SP). This segregation of tasks would ease service providers not to worry about the access management costs and authentication to the IDP and also it receives reliable information about the user even without requiring users to register with the service provider.

FIM supports user single sign-on (SSO) using several of industry standard security protocols and tokens. Moreover, for web service calls it provides web service security [18].

### **3.2.3 Anonymized Data Searching**

This is most important parameter of this research as it will not be using a traditional data de-identifying technique i.e., k-anonymity. Through k-anonymity technique, anonymized data searching parameter of this research is achieved. But as mentioned earlier in chapter 2, there are some serious privacy problems in using k-anonymity. Therefore, an enhanced method for data de-identification could be used in this regard i.e., L-Diversity along with K-Anonymity as shown in figure 3.2.



**Figure 3.2: K-Anonymity Vulnerabilities**

### 3.3 Semantic Validation

The Table 3.4 shows some medical records of an imaginary hospital. It should be kept in mind that the uniquely identifying attributes (such as SSN, name or mobile number etc.) are not published in these medical records. In this scenario, the attributes are divided into two groups:

- Sensitive attributes (contains the medical condition or diagnosis)
- Non-sensitive attributes (contains information such as age, zip code and nationality)

Moreover, assuming these attributes (age, zip code, nationality) to be the quasi-identifiers for this specific dataset. K-anonymity technique is applied on table 3.4 and the results of it are stored in table 3.5 (4-anonymous table). “\*” denotes to be a suppressed value, for instance, “zip code = 1234\*” means zip code is in the range [12340-12349]. Similarly, “age=4\*” means age is in the range of [40-49]. It should be kept in mind that in 4-anonymous table, each record has same quasi-identifier values as the other three medical records.

K-anonymity usage become very common and has gained popularity because of its simplicity and it becomes to be an effective tool for privacy preservation in data publishing [32, 33, 34, 35, 36, 37, 38]. However, the real question still lies that whether k-anonymity really guarantee privacy preservation? The answer to this question is surprisingly “NO”. This answer will be further justified by providing examples of two simple yet very practical attacks on k-anonymous dataset which will give leverage for the attacker to exactly identify that unique

individual. L-Diversity provides defense against these existing vulnerabilities in k-anonymity. Let us first show the two attacks to give the intuition behind the problems with k-anonymity. Now first we will elaborate the two possible attacks on k-anonymity which will give us a direction to the solution for k-anonymity.

**Table 3.4: Medical Records Table**

	<b>Non-Sensitive attributes (Quasi-identifiers)</b>			<b>Sensitive attribute</b>
	<b>Nationality</b>	<b>Age</b>	<b>Zip Code</b>	<b>Diagnosis</b>
1	Pakistani	38	15053	Cardiovascular Disease
2	British	39	15068	Cardiovascular Disease
3	Chinese	31	15068	Flu
4	British	33	15053	Flu
5	Bangladeshi	60	16853	Diabetes
6	Pakistani	65	16853	Cardiovascular Disease
7	British	57	16850	Flu
8	British	59	16850	Flu
9	British	41	15053	Diabetes
10	Bangladeshi	47	15053	Diabetes
11	Chinese	46	15068	Diabetes
12	British	45	15068	Diabetes

**Table 3.5: 4-Anonymous Table**

	Non-Sensitive (Quasi-identifiers)			Sensitive
	Nationality (Suppressed attribute)	Age (Generalized attribute)	Zip Code (Generalized attribute)	Diagnosis
1	*	<40	150**	Cardiovascular Disease
2	*	<40	150**	Cardiovascular Disease
3	*	<40	150**	Flu
4	*	<40	150**	Flu
5	*	≥50	1685*	Diabetes
6	*	≥50	1685*	Cardiovascular Disease
7	*	≥50	1685*	Flu
8	*	≥50	1685*	Flu
9	*	4*	150*	Diabetes
10	*	4*	150*	Diabetes
11	*	4*	150*	Diabetes
12	*	4*	150*	Diabetes

### 3.3.1 Attacks On k-Anonymity

Following attacks are possible on k-anonymity:

#### Homogeneity Attack:

Suppose user1 (Paul) and user2 (David) are unfriendly neighbors. One day user2 gets ill and he was taken to the hospital by an ambulance. User1 was able to see the ambulance and he was wondering about the disease user2 is suffering from. User1 discovers the 4-anonymous table which got published on the hospital website (table 4.2). User1 do know that one of these published records contain user2’s medical record as well. As user1 is user2’s neighbor, he knows that user2 is 41 years old British male who is living in the zip code 15053. This narrows down the records for user1 quite swiftly and he comes to the conclusion that user2’s record lies in number 9, 10, 11

or 12. Now, there is huge problem as records from 9 to 12 all have same medical condition (diabetes), which gives user1 the privilege to know that user2 has diabetes.

### **Observation 1:**

**Groups formed by k-anonymity can compromise information because of lack of diversity in the sensitive attributes.**

This sort of scenario is very common. Considering an example for this type of scenario:

- Dataset records = 101766 unique patient records
- Maximum sensitive attributes in a group = 3 values
- K-anonymization value = 5
- Groups formed =  $\frac{101766}{5} \cong 20353$  groups
- No diversity of 1 complete group (on average) from every 81 groups (Proved by Ashwin Machanavajjhala in their experiments on real dataset)
- Total groups with no diversity =  $\frac{20353}{81} \cong 251$  groups
- Number of individuals whose records would be compromised =  $251 \times 5 = 1256$  records

So, we can say that 1256 patient records are compromised in this scenario via homogeneity attack. Through this example, we got the notion that only k-anonymity alone is not enough, and we should ensure diversity of sensitive attributes in all records of the same group.

### **Background Knowledge Attack:**

Suppose, user1 has a Facebook friend named user3 who is admitted in the same hospital as user2. The patient records are also same as it was in table 3.4. Now user1 has some background knowledge about user3. He knows that user3 is 31 years old, Asian (specifically Chinese), male and he is living in a zip code of 15068. With this much knowledge, user1 has narrowed down user3's health record lies in first group with serial number from 1 to 4. Even with no further knowledge, user1 narrowed it down that user3 either has a cardiovascular disease or flu. Now, it is a widely accepted fact that Chinese people suffered from heart disease percentage is very low. So, user1 can easily conclude with a very high probability that user3 has suffered from flu.

**Observation 2:**

**K-anonymity is not capable to preserve privacy against background knowledge attacks.**

So far, we established the fact that k-anonymity cannot ensure privacy against background knowledge and homogeneity attacks. As we know that both these attacks are very practical and doable, so we need to have a better method than k-anonymity which will be immune against both these attacks.

**Table 3.6: 3-Diversity**

	<b>Non-Sensitive (Quasi-identifiers)</b>			<b>Sensitive</b>
	<b>Nationality</b>	<b>Age</b>	<b>Zip Code</b>	<b>Diagnosis</b>
1	*	≤50	1505*	Cardiovascular Disease
4	*	≤50	1505*	Flu
9	*	≤50	1505*	Diabetes
10	*	≤50	1505*	Diabetes
5	*	>50	1685*	Diabetes
6	*	>50	1685*	Cardiovascular Disease
7	*	>50	1685*	Flu
8	*	>50	1685*	Flu
2	*	≤50	1506*	Cardiovascular Disease
3	*	≤50	1506*	Flu
11	*	≤50	1506*	Diabetes
12	*	≤50	1506*	Diabetes

**3.4 L-Diversity Postulate**

Basically L-Diversity is the amalgamation of two important aspects of privacy, keeping in view that k-anonymity is not neglected:

- Lack of diversity in sensitive attributes of PHRs in a single group.
- Rich & Extensive background knowledge.

## **1. Lack of Diversity:**

For this aspect of privacy, for sensitive attributes nearly all PHRs must have same value. This aspect of privacy can easily be examined through simply counting the different values of sensitive attributes. Therefore, all probable values of sensitive attributes should be in a group with nearly same ratio. Then, we can conclude that such a group is well-represented via  $l$  sensitive values.

## **2. Rich & Extensive Background Knowledge:**

This postulate, no matter how well-represented  $l$  values are in sensitive values, user1 can still reach to the facts with a very high probability using his background knowledge. For instance, let say user1 knows the exact age, nationality and the zip code of user3. Still it will be very difficult to pin point the exact disease user3 is suffering from. Furthermore, let's say that user1 even knows that Chinese people has low percentage to suffer from heart disease, even still he won't be able to say for sure whether user3 suffered from flu or diabetes. So, in a nutshell, we can conclude that user1 will need to have background knowledge of all  $l$  sensitive attributes after which he can narrow it down to reach to the exact diagnosis which is practically impossible. Therefore, when parameter  $l$  is assigned to the sensitive attributes to an  $l$ -diverse table, then the publisher does not need to know what extent of background knowledge the adversary possesses (it can even shield against the adversaries possessing maximum background knowledge).

## **3.5 HIPAA Privacy Rule & Classification of Attributes**

The proposed framework satisfies HIPAA Privacy Rule Section 164.514(a) which provides the standard for de-identification of protected health information. [28]

Standard for PHI de-identification:

Patient health information should not be in a state from which an entity can be specifically identified i.e., direct identifiers. Moreover, in this regard, there should be no reasonable amount of information from which an entity can be identified indirectly i.e., quasi-identifiers.

The following eighteen identifiers should be removed from the microdata if any of them existed:

1. Names
2. ZIP Code
3. All dates that are directly related to an entity, e.g., DoB, date of admission, date of death, date of discharge etc.
4. Phone/Mobile numbers
5. Vehicle identifiers and serial numbers, including license plate numbers
6. Fax numbers
7. Serial numbers and device identifiers
8. E-mail IDs
9. Web Universal Resource Locators (URLs)
10. Social security numbers
11. IP addresses
12. Medical record numbers
13. Biometric identifiers, voice and finger-prints
14. Health plan beneficiary numbers
15. Full-face photographs or any comparable images
16. Account numbers
17. Any other unique identifying number, characteristic, or code
18. Certificate/license numbers

Satisfying either method would demonstrate that a covered entity has met the standard in §164.514(a) above. In this thesis (in experimental setup (section 4.2)), we removed the identifying attributes namely, hospital name, address (home), date of birth, ZIP code and phone numbers. This removal of identifying attributes makes our experimental scenarios meet the HIPAA standard of §164.514(a) [28].

### **3.6 Conclusion**

Former privacy preservation E-Healthcare data security frameworks did not provide strong patient anonymity level, anonymized data searching and successful correlation of PHR for medical research all in a single framework. This framework so far to my knowledge is the only one which provides all these three privacy preservation parameters (i.e. strong patient anonymity level, anonymized data searching and correlating PHR for medical research all in a single framework).



Furthermore, in this framework an enhanced technique for data de-identification is used i.e. L-diversity along with K-Anonymity rather than K-Anonymity alone because of some existing vulnerabilities in this technique which are tackled with the use of L-Diversity. Moreover, in this chapter, the proposed framework is semantically validated. Lastly, HIPAA Privacy Rule is elaborated and shown that the proposed framework meets the requirement of the HIPAA Privacy Rule as well.

## **IMPLEMENTATION & RESULTS**

### **4.1 Introduction**

This chapter gives a proof of the framework proposed in chapter 3 by validating it through case scenarios and more importantly experimentally it will provide evidence of better results of L-Diversity and K-anonymity instead of using K-anonymity alone. The tool for experimental implementation used is ARX Anonymization Tool [29]. The case scenario is quite practical and doable in real life. A simple microdata is first shown on which k-anonymization is applied. The results from k-anonymity are shown elaborating that it does not provide the required privacy. Then L-diversity is applied along with K-Anonymity which meets the essential privacy requirements and also it shows that it is immune to the attacks possible on K-Anonymity.

### **4.2 Experimental Setup**

ARX Anonymization Tool is used for this research. Different case scenarios are developed in ARX and results are shown in the end for the comparative analysis between k-anonymity and L-Diversity. For the ease of our data analysis we combine two different datasets in order to meet the HIPAA Privacy Rule in which we can remove identifying attributes as well as mark sensitive and quasi-identifier attributes. Total number of EMRs dataset used for this research are 101,766 [30]. In this dataset, there are total 11 attributes namely, hospital name, address, state, phone number, ZIP code, age, gender, race, weight, insulin and diabetesMed.

To meet the HIPAA Privacy requirement, we have totally suppressed the identifying attributes such as hospital name, address, phone number. Insulin and diabetesMed is considered as the sensitive attributes in this experimental approach and the rest of the attributes are considered as quasi-identifier attributes. Weight attribute however is found fully suppressed by the publishers who published this dataset on [30].

#### **4.2.1 ARX Anonymization Tool**

ARX is an open source tool used for privacy preservation through data anonymization. It supports both k-anonymity and L-Diversity privacy models. We will be doing a comparative

analysis between both these privacy models on ARX. Three types re-identification risk factors are used in it, namely [20]:

- The Prosecutor Scenario
- The Journalist Scenario
- The Marketer Scenario

### **1. The Prosecutor Scenario:**

In this scenario, attacker possess background knowledge of the target entity that is present in the dataset.

### **2. The Journalist Scenario:**

This scenario is totally opposite of prosecutor scenario. In this scenario, attacker has no background knowledge of target entity that is present in dataset or not.

### **3. The Marketer Scenario:**

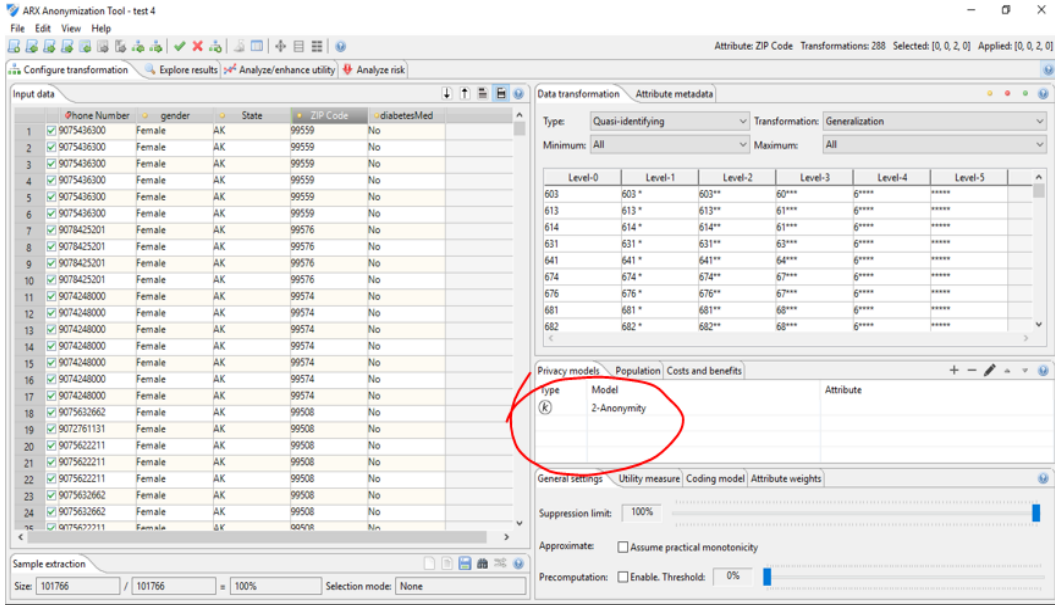
In this scenario, attacker is not interested in the specific entity's identification, rather he is interested in successfully identifying a specific percentage of records in the dataset. This scenario is very important for our research point of view as results are based on this specific scenario and the reason for the selection of this criteria is that it is very common compared to other scenarios. The attackers are more interested in the more amount of EMRs compromised rather than targeting a specific identity which is a very rare.

## **4.3 Case Scenarios**

Different K-Anonymity and L-Diversity values are chosen for these case scenarios and percentage of records at risk are given at the end of each scenario in order to simplify the comparison between both these techniques.

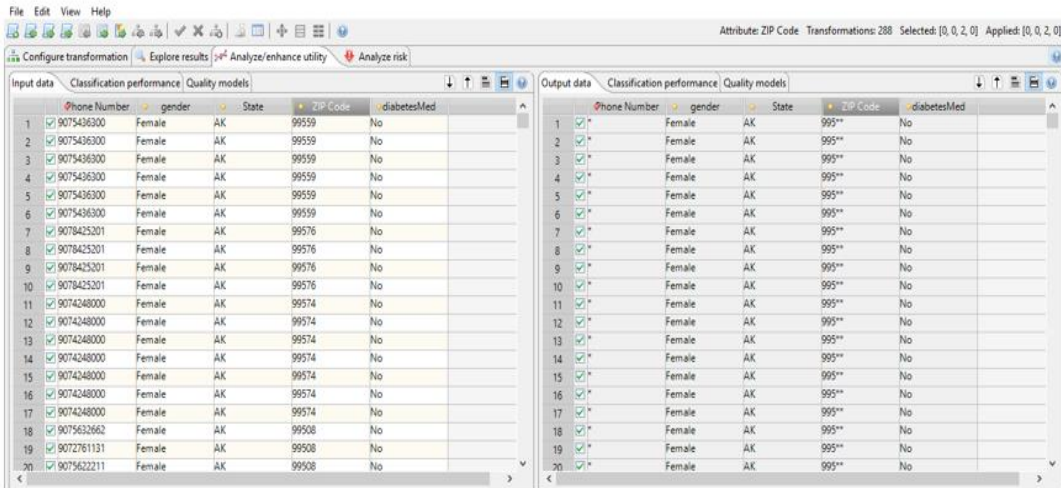
### **4.3.1 2-Anonymity**

In this scenario, we chose value for k-anonymity = 2 ( $k = 2$ ).  
L-Diversity not implemented in this scenario.



**Figure 4.1: 2-Anonymity**

Figure 4.1 shows 2-Anonymity implementation in ARX anonymization tool. In the following figure 4.2, shows the input and output EMRs after implementing 2-Anonymity.



**Figure 4.2: 2-Anonymity Input and Output EMRs**

**Input**

This is risk analysis percentage before applying anonymization. (NOTE: We are considering the marketer attacker model as explained in section 4.2.1. in the marketer scenario).

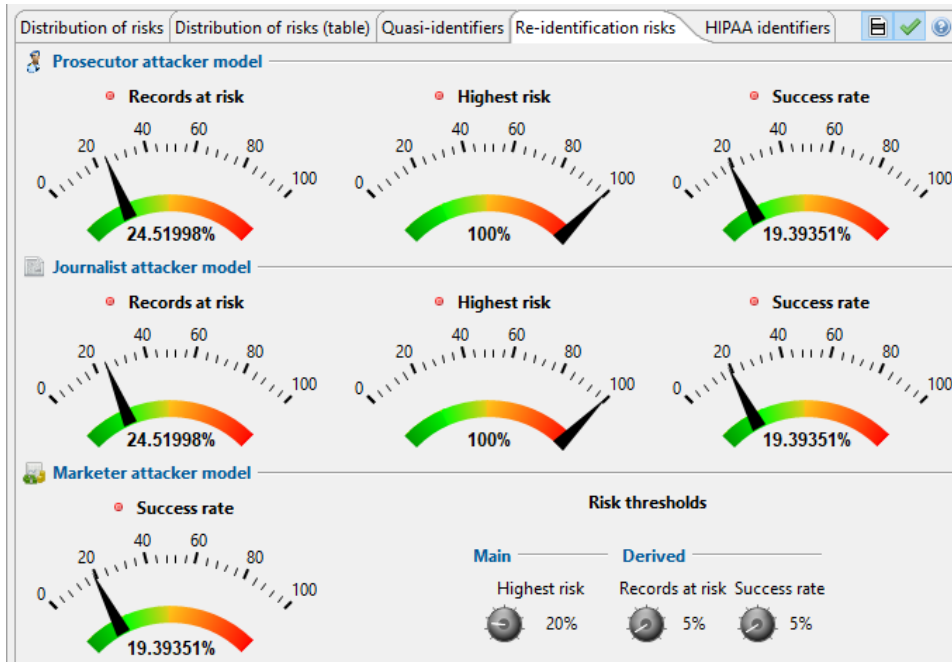


Figure 4.3: Risk Analysis before 2-Anonymity

## Output

This is risk analysis percentage after applying anonymization. (NOTE: We are considering the marketer attacker model as explained in section 4.2.1. in the marketer scenario)

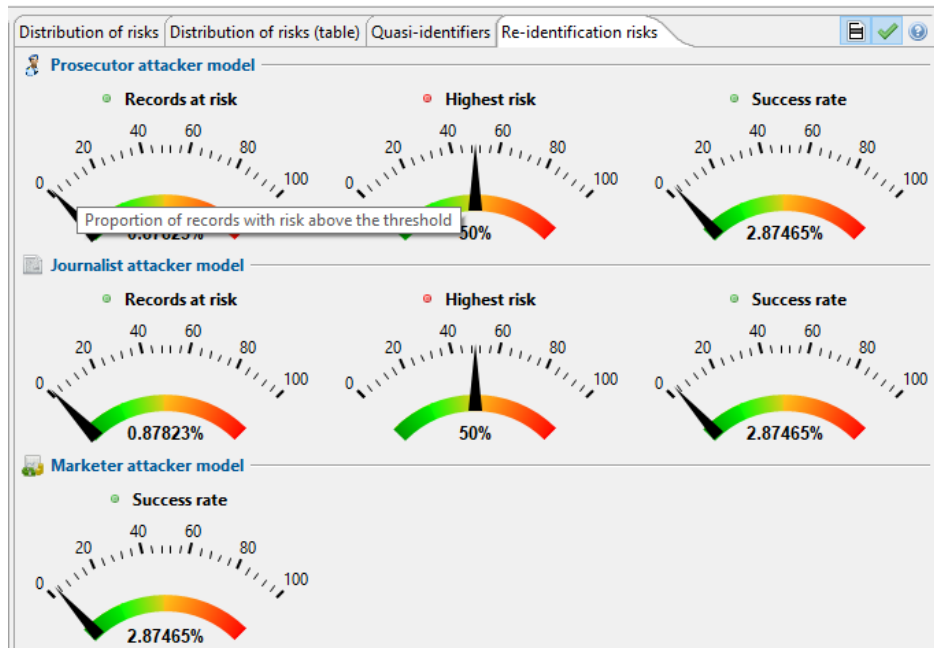


Figure 4.4: Risk Analysis after 2-Anonymity

From this scenario onwards, table 4.1 is made in which risk analysis values are shown to get a better insight instead of using the above figures for every scenario. Values of marketer attacker model are used.

**Table 4.1: Percentage of Risk Analysis on different Anonymization Values**

Scenarios (Sr. No.)	K-Anonymity	L-Diversity	Input (Percentage of records at risk before anonymization)	Output (Percentage of records at risk after anonymization)	Explanation in section
1.	2	-	19.3	2.87	4.4.1
2.	3	-	19.3	1.14	
3.	5	-	19.3	1.03	
4.	5	2	19.3	1.19	4.4.2
5.	5	3	19.3	0.308	
6.	5	4	19.3	0.29	

#### 4.4 Results

Results on these six scenarios shown above are elaborated in this section graphically using bar charts for better understanding.

##### 4.4.1 K-Anonymization Plots

###### Inputs

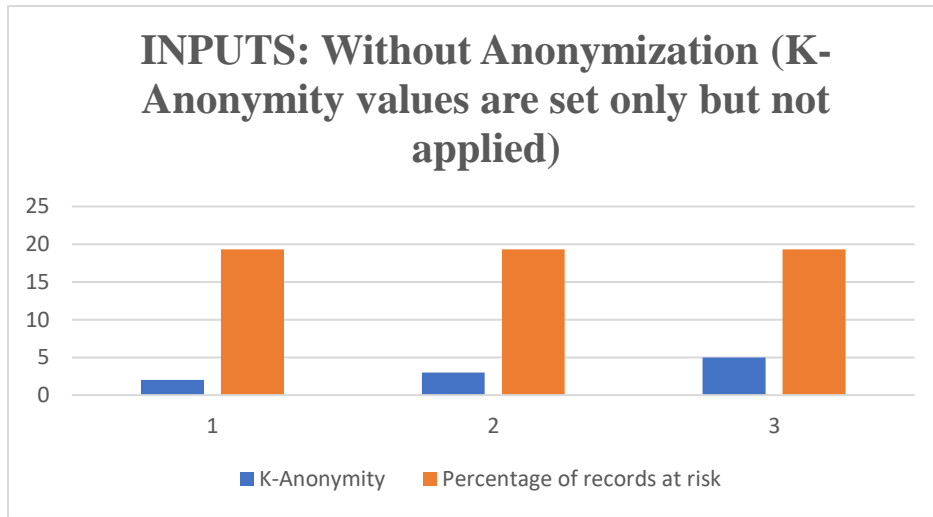
**NOTE:** K-Anonymity values are set only but not applied on dataset for anonymization yet.

**Table 4.2: K-Anonymization values w.r.t. Percentage of Records at Risk before anonymization**

K-Anonymity	Percentage of records at risk
2	19.3
3	19.3
5	19.3

Before applying any anonymization techniques, percentage of records at risk are 19.3% on all values of k used in the experimental scenario. Total number of records used are 101766. Records at risk before applying k-anonymization are:

$$19.3\% \text{ of } 101766 = \frac{19.3}{100} \times 101766 = 19640 \text{ Records}$$



**Figure 4.5: K-Anonymization values w.r.t. Percentage of Records at Risk before Anonymization**

Graphical view also illustrates here depicting results of table 3 that without applying anonymization technique on dataset, percentage of records at risk are 19.3%.

## Output

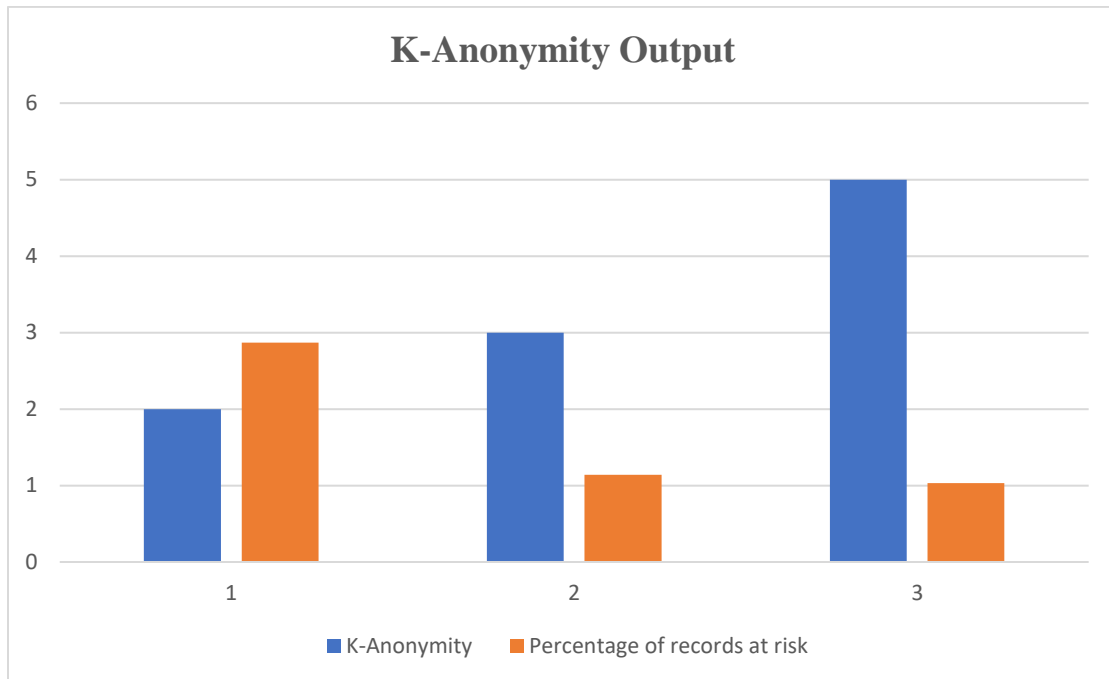
**Table 4.3: K-Anonymization values w.r.t. Percentage of Records at Risk after anonymization**

K-Anonymity	Percentage of records at risk
2	2.87
3	1.14
5	1.03

After applying K-Anonymity technique, percentage of records at risk are shown below on all values of k used in the experimental scenario. Total number of records used are 101766. Records at risk after applying k-anonymization are:

➤ Records at risk after K-Anonymization

- 2-Anonymization = 2.87 % of 101766 =  $\frac{2.87}{100} \times 101766 = 2920$  Records
- 3-Anonymization = 1.14 % of 101766 =  $\frac{1.14}{100} \times 101766 = 1160$  Records
- 5-Anonymization = 1.03 % of 101766 =  $\frac{1.03}{100} \times 101766 = 1048$  Records



**Figure 4.6: K-Anonymization values w.r.t. Percentage of Records at Risk after Anonymization**

Graphical view also illustrates here depicting results of table 4 that after applying anonymization technique on dataset, percentage of records at risk are as follows. With value of k equal to 2, percentage of records at risk are 2.87%. With value of k equal to 3, percentage of records at risk are 1.14% and with value of k equal to 5, percentage of records at risk are 1.03%.

#### 4.4.2 K-Anonymity with L-Diversity Plots

**Input:**

**NOTE:** K-Anonymity and L-Diversity values are set only but not applied on dataset for anonymization yet.

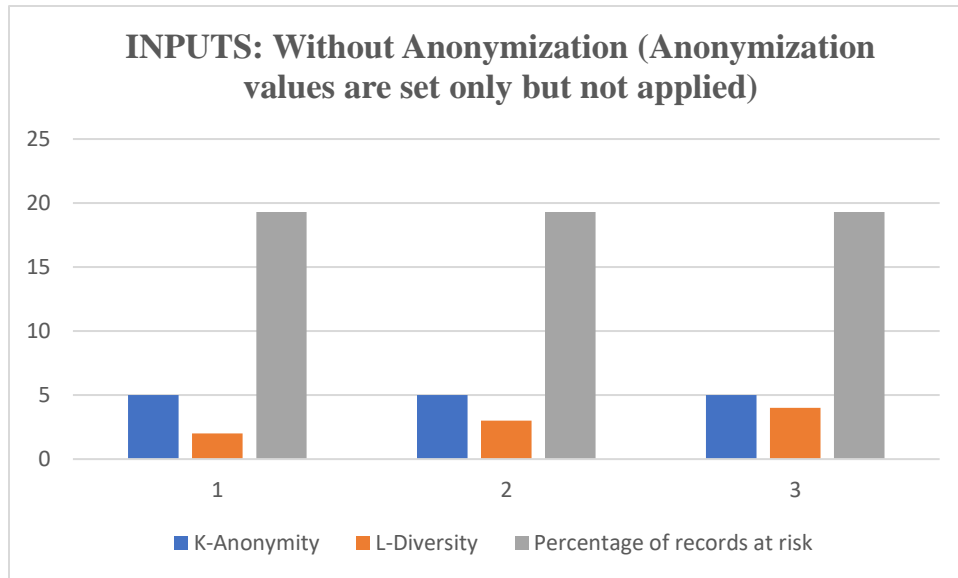


**Table 4.4: K-Anonymization and L-Diversity values w.r.t. Percentage of Records at Risk before Anonymization**

K-Anonymity	5	5	5
L-Diversity	2	3	4
Percentage of records at risk	19.3	19.3	19.3

Before applying any anonymization techniques, percentage of records at risk are 19.3% on all values of K and L used in the experimental scenario. Total number of records used are 101766. Records at risk before applying K-Anonymization and L-Diversity are:

$$19.3\% \text{ of } 101766 = \frac{19.3}{100} \times 101766 = 19640 \text{ Records}$$



**Figure 4.7: K-Anonymization & L-Diversity values w.r.t. Percentage of Records at Risk before Anonymization**

Graphical view also illustrates here depicting results of table 5 that without applying anonymization techniques of K-Anonymity and L-Diversity on dataset, percentage of records at risk are 19.3%.

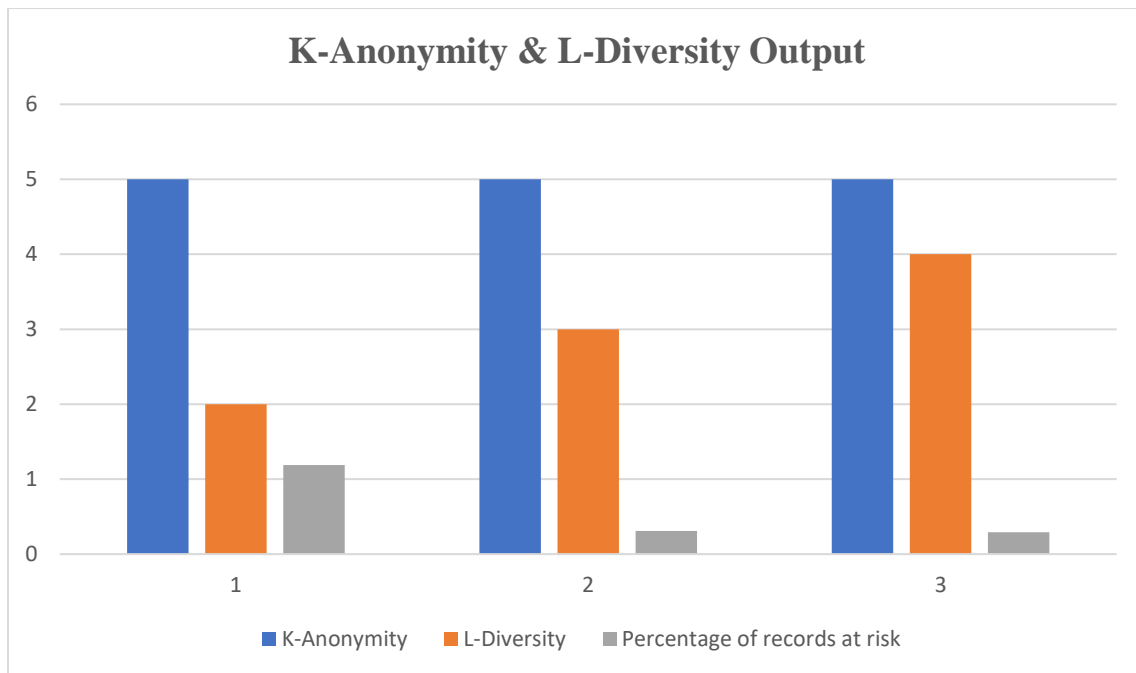
## Output

**Table 4.5: K-Anonymization and L-Diversity values w.r.t. Percentage of Records at Risk after Anonymization**

K-Anonymity	5	5	5
L-Diversity	2	3	4
Percentage of records at risk	1.19	0.308	0.29

After applying K-Anonymity and L-Diversity techniques, percentage of records at risk are shown below on all values of K and L used in the experimental scenario. Total number of records used are 101766. Records at risk after applying K-Anonymization and L-Diversity are:

- Records at risk after 5-Anonymization & L-Diversity
  - 2-Diversity = 1.19 % of 101766 =  $\frac{1.19}{100} \times 101766 = 1211$  Records
  - 3-Diversity = 0.308 % of 101766 =  $\frac{0.308}{100} \times 101766 = 313$  Records
  - 4-Diversity = 0.29 % of 101766 =  $\frac{0.29}{100} \times 101766 = 295$  Records



**Figure 4.8: K-Anonymization & L-Diversity values w.r.t. Percentage of Records at Risk**

Graphical view also illustrates here depicting results of table 6 that after applying anonymization techniques of K-Anonymization and L-Diversity on dataset, percentage of records at risk are as follows. Value of K is set to 5 in all these three scenarios. With value of L equal to 2, percentage of records at risk are 1.19%. With value of L equal to 3, percentage of records at risk are 0.308% and with value of L equal to 4, percentage of records at risk are 0.29%.

#### 4.5 For Verification of Results

Verification of an experimental process is a vital step in order to ensure the accuracy of results. For verification process, another dataset of 1074 EMRs was chosen. Similar case scenarios with same values of K-Anonymity and L-Diversity are developed so that percentage difference could be calculated at the end from each case scenario.

##### 4.5.1 2-Anonymity

In this scenario, value for k-anonymity = 2 (k = 2).

L-Diversity not implemented in this scenario.

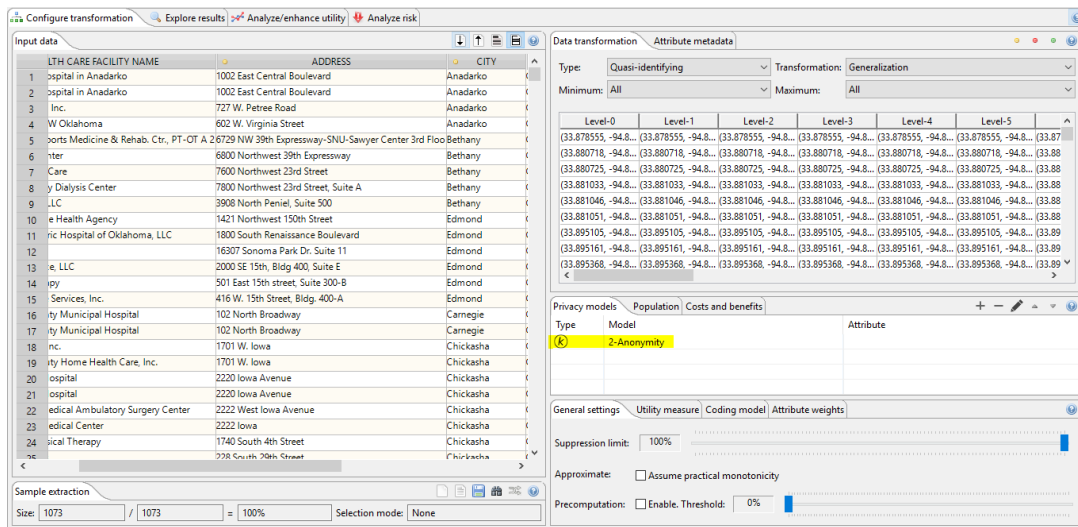


Figure 4.9: 2-Anonymity

Figure 4.9 shows 2-Anonymity implementation on ARX anonymization tool.

#### Input

20.9% is risk analysis percentage before applying any anonymization technique which was previously 19.3% on the original dataset.

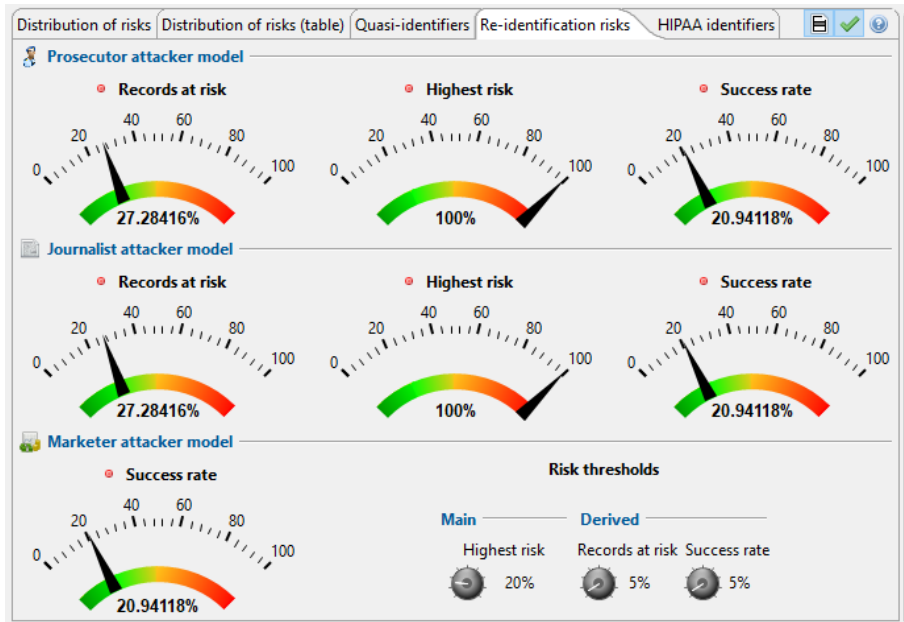


Figure 4.10: Risk Analysis before applying any Anonymization Technique

## Output

5.65% is risk analysis percentage after applying 2-Anonymity which was previously 2.87% on the original dataset.

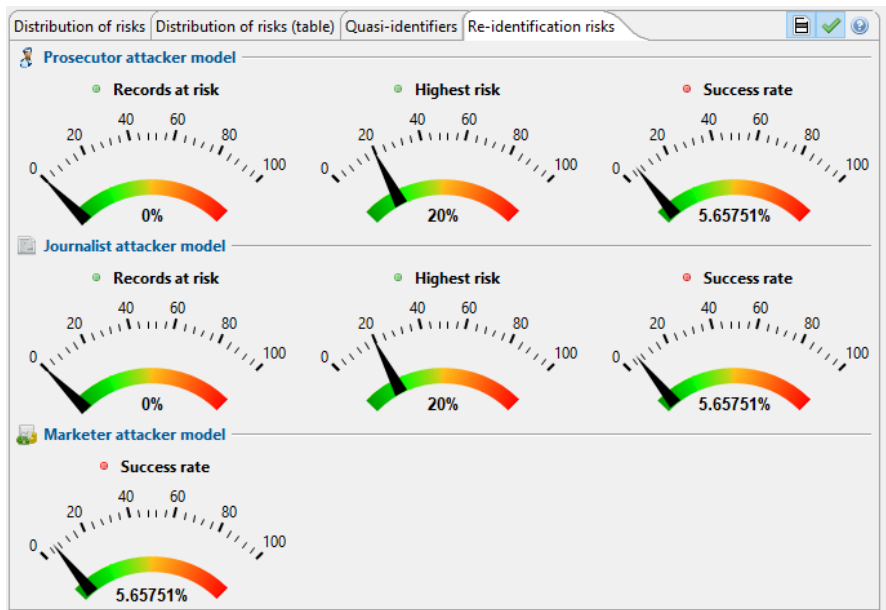


Figure 4.11: Risk Analysis after 2-Anonymity

From this scenario onwards, table 4.6 is made in which risk analysis values are shown to get a better insight instead of using the above figures for every scenario. Values of marketer attacker model are used.


**Table 4.6: Percentage of Risk Analysis on different Anonymization Values**

Scenarios (Sr. No.)	K-Anonymity	L-Diversity	Input (Percentage of records at risk before anonymization)	Output (Percentage of records at risk after anonymization)
1.	2	-	20.9	5.65
2.	3	-	20.9	5.15
3.	5	-	20.9	2.83
4.	5	2	20.9	1.19
5.	5	3	20.9	1.13
6.	5	4	20.9	0.64


For the verification of results, percentage risk analysis of both datasets is compared. In table 4.7, input EMRs from both datasets is shown with a percentage difference of 1.6%.

**Table 4.7: Percentage of Input Risk Analysis of both Datasets**

Scenarios (Sr. No.)	K-Anonymity	L-Diversity	Input (Percentage of records at risk before anonymization)	Output (Percentage of records at risk after anonymization)
1.	2	-	19.3	20.9
2.	3	-	19.3	20.9
3.	5	-	19.3	20.9
4.	5	2	19.3	20.9
5.	5	3	19.3	20.9
6.	5	4	19.3	20.9



Original Dataset



2<sup>nd</sup> Dataset

Output EMRs from both datasets is shown in table 4.8 along with percentage difference of risk analysis of both datasets.

**Table 4.8: Percentage of Output Risk Analysis of both Datasets**

Scenarios (Sr. No.)	K-Anonymity	L-Diversity	Output (Percentage of records at risk after anonymization)	Output (Percentage of records at risk after anonymization)	Percentage Difference
1.	2	-	2.87	5.65	2.78
2.	3	-	1.14	5.15	4.01
3.	5	-	1.03	2.83	1.8
4.	5	2	1.19	1.19	0
5.	5	3	0.308	1.13	0.822
6.	5	4	0.29	0.64	0.35

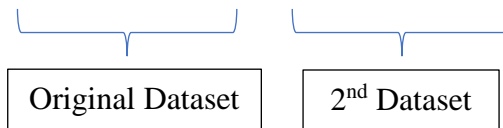


Table 4.7 and table 4.8 provides clear similarities between both datasets with slight percentage differences because of the sizes of datasets. From table 4.8, a decreasing trend in percentage of records at risk is observed as K-Anonymity and L-Diversity values are increased. From the percentage of records at risk values of second dataset, these values are found to be quite near to the original results. From this observation, we can conclude that our results from original dataset were more accurate due to much higher number of records from second dataset.

#### 4.6 Conclusion

From the results obtained in this chapter, we concluded that K-anonymity alone is not only vulnerable, but it also compromises a huge number of EMRs. However, if we use L-Diversity along with K-Anonymity, it not only makes it secure against the attacks on K-Anonymity but also it reduces the percentage of EMRs at risk with considerable number. Furthermore, results are also verified by using another dataset which produces similar percentage results and a decreasing trend of percentage of records at risk is observed with increasing values of K-Anonymity and L-

Diversity. Moreover, the accuracy of original dataset is also verified because of the similarity in percentages from both datasets.

## **CONCLUSION & FUTURE DIRECTIONS**

### **5.1 Conclusion**

Three privacy preservation parameters are used in this research i.e., (patient anonymity level, correlating PHR for medical research and anonymized data searching). Most of the work in this research is done on the anonymized data searching part where previously k-anonymity method was used for data de-identification. The k-anonymization technique discussed in detail in chapter 3 which is quite vulnerable due to which L-Diversity is used along with K-Anonymity to counter any possible attacks against K-Anonymity. Moreover, in chapter 4, we have elaborated experimentally that using L-Diversity along with K-Anonymity reduces a huge number of records at risk which could not be achieved with K-Anonymity alone.

In this research, a privacy preservation framework is made which has all the three considered parameters incorporated in it. Moreover, with a novel change in this framework by using L-Diversity along with K-Anonymity instead of K-Anonymity alone has secured huge number of records which were previously at risk. Moreover, this framework is also compliant with HIPAA privacy rule as the identifiers used are mentioned in HIPAA privacy rule section 164.514(a) [28]. From the experimental analysis, huge number of records were made secured. A sample of 101766 medical records were used in which with K-Anonymity technique on value set at  $k=5$ , only 1048 records were at risk but when L-Diversity is used along with K-Anonymity with value set at  $k=5$  and  $l=4$ , only 295 records were at risk out of 101766 records which means that this hybrid technique was able to secure 99.71% of medical records.

### **5.2 Future Directions**

There is still room for further research in this field. As only one parameter is selected (anonymized data searching) and improved its security. The other two parameters (patient anonymity level and correlation of PHR for medical research) has still room for improvement.

For the patient anonymity level, already AES-128, AES-192 and AES-256 are used which provides maximum security so far but due to continuous advancements in quantum computing,



they are also not much far to be broken. To counter against this, quantum cryptographic algorithms should be put in use which will help in security of privacy preservation.

For the second component which is correlation of PHR for medical research, there is also a vulnerability which could be improved. The federated identity management is used for the fulfillment of this component which delegate authentication to an external identity provider. The vulnerability is that when a client requests for the identity security token, its identity is also attached to its request for identity security token from Identity Provider (IdP). This could be sniffed or intercepted by an attacker who could be eavesdropping and could even masquerade the identity of the client or worse the identity provider.

### **5.3 Summary**

This chapter has concluded the research work by providing a brief overview of the research conducted along with a brief explanation of results in experimental analysis of chapter 4. Furthermore, it has also provided an insight of future direction for the researchers in this field.

## REFERENCES

- [1] Yang, Ji-Jiang, Jian-Qiang Li, and Yu Niu. "A hybrid solution for privacy preserving medical data sharing in the cloud environment." *Future Generation Computer Systems* 43 (2015): 74-86.
- [2] Agrawal, Rakesh, and Christopher Johnson. "Securing electronic health records without impeding the flow of information." *International journal of medical informatics* 76.5 (2007): 471-479.
- [3] Alhaqbani, Bandar, and Colin Fidge. "Privacy-preserving electronic health record linkage using pseudonym identifiers." *e-health Networking, Applications and Services, 2008. HealthCom 2008. 10th International Conference on. IEEE, 2008.*
- [4] Pommerening, Klaus & Reng, Michael & Debold, Peter & Semler, Sebastian C.. (2005). Pseudonymization in medical research - the generic data protection concept of the TMF. *GMS Medizinische Informatik, Biometrie und Epidemiologie*. 1.
- [5] B. Riedl, V. Grascher, S. Fenz and T. Neubauer, "Pseudonymization for improving the Privacy in E-Health Applications," *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), Waikoloa, HI, 2008,* pp. 255-255.
- [6] J. Wang, Y. Zhao, S. Jiang and J. Le, "Providing privacy preserving in Cloud computing," *3rd International Conference on Human System Interaction, Rzeszow, 2010,* pp. 472-475.
- [7] Aamot H., Dominik Kohl C., Richter D. and Knaup-Gregori P., "Pseudonymization of patient identifiers for translational research." *BMC medical informatics and decision making* 13.1 (2013): 75.
- [8] Waqar, A., Raza, A., Abbas, H., Khan, M.K. A framework for preservation of cloud users' data privacy using dynamic reconstruction of metadata. *Journal of Network and Computer Applications, Volume 36, Issue 1, Pages 235-248, January 2013.*

- [9] Win KT, Susilo W, Mu Y. Personal health record systems and their security protection. *J Med Syst* 2006;30(4):309–15.
- [10] S. Narayan, M. Gagné, and R. Safavi-Naini, “Privacy preserving EHR system using attribute-based infrastructure,” *The ACM Cloud Computing Security Workshop, (CCSW ’10)*, October 2010, pp.47–52.
- [11] Shortliffe, H.E. and Blois, M.S., *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, 2000.
- [12] Barua M., Liang X., Lu R. and Shen X., *ESPAC: Enabling Security and Patient-centric Access Control for eHealth in cloud computing*, 2011.
- [13] Kahnamoui, *Rationing critical care beds: A systematic review*, 2004.
- [14] Johan G. Beun, *Electronic healthcare record; a way to empower the patient*, 2003.
- [15] EU Directive 95/46/EC, NHS Code of Practice, National survey of British public's views on use of identifiable medical data by the National Cancer Registry, 2003.
- [16] Alan F. Westin. *Privacy and Freedom*. The Bodley Head Ltd, 1970.
- [17] Ronald Hes and John Borking. *Privacy enhancing technologies: the path to anonymity (revised edition)*. The Dutch Data Protection Authority, September 1998.
- [18] IBM Developer Works, Federated identity management concepts, <https://www.ibm.com/developerworks/library/se-jitp/>.
- [19] Rindfleisch, T.C., *Privacy, information technology, and health care*, 1997.
- [20] Yuce, Baris & Mourshed, Monjur & Rana, Omer & Rezgui, Yacine. (2016). *Preserving Prosumer Privacy in a District Level Smart Grid*.
- [21] Margaret Rouse, *Advanced Encryption Standard (AES)*, Available at: <http://searchsecurity.techtarget.com/definition/Advanced-Encryption-Standard>
- [22] Yoo, S., Shin, M., & Lee, D. (2012). *An Approach to Reducing Information Loss and Achieving Diversity of Sensitive Attributes in k-anonymity Methods*.

Interactive Journal of Medical Research, 1(2), e14.  
<http://doi.org/10.2196/ijmr.2140>.

- [23] English V, Mussell R, Sheather J, et al Ethics briefings Journal of Medical Ethics 2004;30:517-518.
- [24] Stan. Tech. L. Rev. 1, Fair Information Practices and the Architecture of Privacy, 2001.
- [25] Waegemann CP. The five levels of electronic health records. MD Comput. 1996 May-Jun;13(3) 199-203. PMID: 8935995.
- [26] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In Proc. of the 17th ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, 188, 1998.
- [27] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramaniam. L-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intl. Conf. Data Engg. (ICDE), page 24, 2006.
- [28] Methods for De-identification of PHI | HHS.gov “<https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard>”.
- [29] ARX Data Anonymization Tool “<https://arx.deidentifier.org/>”.
- [30] Health IT Dashboard “<https://dashboard.healthit.gov/datadashboard/data.php>”.
- [31] L. Sweeney. Uniqueness of simple demographics in the u.s. population. Technical report, Carnegie Mellon University, 2000.
- [32] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–570, 2002.
- [33] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. k-anonymity: Algorithms and hardness. Technical report, Stanford University, 2004.
- [34] R. J. Bayardo and R. Agrawal. Data privacy through optimal kanonymization. In ICDE-2005, 2005.

- [35] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: Efficient fulldomain k-anonymity. In SIGMOD, 2005.
- [36] A. Meyerson and R. Williams. On the complexity of optimal kanonymity. In PODS, 2004.
- [37] P. Samarati. Protecting respondents' identities in microdata release. In IEEE Transactions on Knowledge and Data Engineering, 2001.
- [38] S. Zhong, Z. Yang, and R. N. Wright. Privacy-enhancing kanonymization of customer data. In PODS, 2005.
- [39] Ji-Jiang Yang, Jianqiang Li, Jacob Mulder, Yongcai Wang, Shi Chen, Hong Wu, Qing Wang, Hui Pan, "Emerging information technologies for enhanced healthcare", Computers in Industry, Volume 69, 2015, Pages 3-11.
- [40] Martin Henze, Lars Hermerschmidt, Daniel Kerpen, Roger Häußling, Bernhard Rumpe, Klaus Wehrle, "A comprehensive approach to privacy in the cloud-based Internet of Things", Future Generation Computer Systems, Volume 56, 2016, Pages 701-718.
- [41] David F. Ferraiolo, Ravi Sandhu, Serban Gavrilă, D. Richard Kuhn, and Ramaswamy Chandramouli. 2001. Proposed NIST standard for role-based access control. ACM Trans. Inf. Syst. Secur. 4, 3 (August 2001), 224-274.
- [42] D. Slamanig, C. Stingsl, The degree of privacy in web-based electronic health records, in: 4th European Conference of the International Federation for Medical and Biological Engineering, Springer, 2009, pp. 974–977.
- [43] Shamir A. How to share a secret. Commun ACM 1979;22(11):612–3.
- [44] J. Fingberg, M. Hansen, M. Hansen, H. Krasemann, L. Lo Iacono, T. Probst, and J. Wright: Integrating Data Custodians in eHealth Grids - A Digest of Security and Privacy Aspects, In: Christian Hochberger, and Rüdiger Liskowsky (Eds.), Informatik 2006 - Informatik für Menschen, Lecture Notes in Informatics (LNI) vol. P-93, pp. 695-701, 2006.

- [45] Bickford, Julia & Nisker, Jeff. (2014). Tensions Between Anonymity and Thick Description When "Studying Up" in Genetics Research. *Qualitative health research*. 25. . 10.1177/1049732314552194.
- [46] A. Shah, H. Abbas, W. Iqbal and R. Latif, "Enhancing E-Healthcare Privacy Preservation Framework through L-Diversity," 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC), Limassol, Cyprus, 2018, pp. 394-399.