

**PRIVACY-AWARE CLOUD-BASED ARCHITECTURE FOR SHARING
HEALTHCARE INFORMATION**

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY IN
FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF
PHILOSOPHY

Supervisors

Professor Jairo Gutierrez

Dr. William Liu

June 2020

By

Fadi Jamil Alhaddadin

School of Engineering, Computer, and Mathematical Sciences

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning, except where explicitly defined in the acknowledgments.

Fadi Jamil Alhaddadin

Dedication

This thesis work is dedicated to my parents Jamil Alhaddadin and Lydia Ayoub who have always loved me unconditionally and immensely inspired me during my Ph.D. journey. I am truly thankful for having you in my life.

This work is also dedicated to my lovely wife Celina Abo who has never left my side all the way through. It is a genuine pleasure to express my deep sense of gratitude to you for your practical and emotional support. I love you and wholeheartedly thank you for being in my life.

I also dedicate this work to my two-and-a-half-year-old son Jaime, who probably thought his father would be studying for the rest of his life. He will now get more time to sing the Baby Shark song with his father.

Acknowledgments

I would like to acknowledge the people who have helped me during this strenuous yet extremely rewarding journey towards the completion of this thesis. First and foremost, I give my sincere thanks to my primary research supervisor Professor Jairo Gutierrez for his constant and valuable guidance throughout the research journey. Also, my second supervisor Dr. William Liu whose feedback has always been valuable to the outcome of the research.

Nobody has been more important to me in the pursuit of this thesis than my family members who have always believed in my ability to succeed. With boundless love and appreciation, I thank my parents who showered me with prayers and love throughout my academic journey. I also thank my caring and supportive wife, Celina has been extremely supportive to me throughout the entire journey, she has made countless sacrifices to help me complete this thesis. I also thank my wonderful son Jaime, who always provided endless inspiration. Also, great thanks go to my parents-in-law Yacoub Abo and Margo Haddad for their prayers and words of encouragement.

I would also like to thank my colleagues and friends who have been instrumental in supporting and motivating me during this journey. To name some of them, I thank Dr. Eghbal Ghazi Zadeh for his support when I needed it. Special thank you goes to Dr. Omar Ababneh who has always been a source of genuine advice and motivation for me during the conduction of this research. I also thank my friends Sami Jamhour and Elias Abo for motivating me and sharing the good times.

Last but not least, the assistance of AUT administrators, IT services, the Library, and in particular, AUT postgraduate office are also acknowledged with gratitude.

Abstract

The advancement in the field of information and communications technology has led to generating a significant amount of information in various fields and domains. The healthcare industry like other industries has generated large amounts of information-driven by record-keeping, compliance, regulatory requirements, and of course the patient care. This has resulted in a large amount of data that has volume, enormous velocity, and a vast variety which makes hospitals today tend to implement electronic health record systems (EHR systems).

Patient health-related information generates special value when it is shared and collaboratively used among different parties involved in the healthcare domain. Several interviewed experts consider immediate access to previously generated medical records during healthcare service delivery as highly important. The use of collected data is a valuable source for analysis that benefits both medical research and practice. Information systems in the healthcare domain are required to collaborate by exchanging information for medical care purposes.

In the healthcare domain, patients usually acquire medical care from various caregivers such as hospitals, pharmacies, laboratories, school clinics, public health places, etc., and as a consequence, information collected about patients is stored in different locations, making it difficult to access when a holistic picture of the patient's health is required for medical treatment purposes. The challenge for exchanging information among heterogeneous systems is related to two aspects namely lack of interoperability and information privacy-related concerns. To realize the full potential of collected medical data, healthcare information systems and products are required to share information seamlessly among each other, but unfortunately, the vast majority of medical devices, electronic health records, and other information technology systems lack interoperability. Privacy is another challenge that hinders the share of information among different parties in the healthcare sector. The privacy-related regulations are considered one of the biggest challenges to healthcare data sharing. Such regulations prohibit the transmission of personal health information among collaborating organizations impeding research and reducing the utility of the datasets.

Cloud computing matches the need of healthcare information sharing directly to various healthcare-related parties over the internet, regardless of their locations and the amount of data being shared. However, the adoption of cloud computing in the healthcare domain requires solving several issues and information privacy is a major one.

This thesis aimed to identify the desired characteristics of healthcare information systems, and further propose a solution for adopting the cloud technology for sharing healthcare information in a privacy-preserving manner. The research was conducted in a multi-methodological approach underpinned by the Design Science research methodology. A case study method was followed for identifying the characteristics required for healthcare information systems. Six healthcare-related institutions participated in the research from which medical practitioners were interviewed.

A cloud architecture design for the healthcare information system was proposed. The proposed architecture enables for storing and sharing patient information for both; medical treatment and research purposes in a privacy-preserving manner. Patients information in the proposed architecture is divided into four categories identified in the case study data analysis. User identity management protocol (U-IDM) is employed for controlling the access to patients' information that is stored in the cloud, and patients are granted with means of control over who can access their information. Further, the proposed architecture enables for sharing healthcare information for research purposes in a privacy-preserving manner; it performs many anonymization operations on patients' information to preserve the privacy of the information when it is aggregated and used for research purposes.

A scenario-based instantiation was developed for validating the proposed architecture in terms of sharing patient information in a privacy-preserving manner. The instantiation showed that the proposed architecture allows for sharing healthcare information without compromising the privacy of individual patients concerning the privacy policies and regulations relating to healthcare information.

Table of Contents

Declaration.....	ii
Dedication.....	iii
Acknowledgements.....	iv
Abstract.....	v
List of Figures.....	xi
List of Tables.....	xii
List of Abbreviations.....	xiv
Chapter 1: Introduction.....	1
1.1 Background.....	3
1.1.1 Interoperability Challenge.....	5
1.1.2 Privacy Challenge.....	7
1.2 Research Problem.....	8
1.3 Research Motivation and Significance.....	11
1.4 Research Questions.....	11
1.5 Thesis Publications.....	12
1.6 Thesis Structure.....	13
Chapter 2: Literature Review.....	15
2.1 Interoperability.....	16
2.2 Anonymization.....	22
2.2.1 <i>K</i> -anonymity.....	24
2.2.2 <i>l</i> -diversity.....	25
2.2.3 <i>t</i> -closeness.....	26
2.2.4 Information privacy threats.....	27
2.3 Cloud Computing Solution.....	29
2.3.1 Barriers to adopting cloud computing.....	32
2.3.2 Cloud computing for healthcare.....	33

2.3.3 Privacy-preservation approaches	37
2.4 Summary	42
Chapter 3: Research Methodology and Design	44
3.1 Research Design.....	45
3.2 Case Study Approach	48
3.2.1 Research Questions.....	49
3.2.2 Research Proposition	49
3.2.3 Unit of Analysis.....	49
3.2.4 The logic that links data to propositions.....	51
3.2.5 Criteria for interpreting findings.....	51
3.3 Data Analysis	52
3.4 Summary	56
Chapter 4: Data Collection and Findings.....	58
4.1 Data Gathering	58
4.1.1 Interview Protocol	60
4.1.2 Data Transcribing and Preparation	60
4.1.3 Organizing the Data.....	60
4.2 Discussion	62
4.2.1 Background.....	62
4.2.2 Interview questions.....	63
4.3 Study Findings	65
4.3.1 Information Needs	65
4.3.2 Desired System Characteristics	70
4.4 Summary	74
Chapter 5: Proposed Cloud Architectural Design.....	75
5.1 Storing and Sharing Information.....	78
5.1.1 Structuring Patients Information	78

5.1.2 Searchable Symmetric Encryption (SSE).....	81
5.1.3 Architectural Design and Components.....	86
5.2 Information for research purposes.....	92
5.2.1 Privacy preservation strategies	92
5.2.2 Architectural Design and Components.....	102
5.2.3 Storing patients records on the Research Portal Server.....	103
5.2.4 Releasing datasets for research purposes.....	105
5.3 Summary	109
Chapter 6: System Instantiation	110
6.1 Storing patient information on the cloud - Scenario	110
6.2 Accessing stored patient information – Scenario	115
6.2.1 Protocol to access information stored in the cloud.....	116
6.2.2 Updating patient information.....	120
6.2.3 Security Analysis of the Proposed system design	121
6.3 Architecture Implementation.....	125
6.3.1 AWS services used	126
6.3.2 Implementation Objectives.....	128
6.3.3 Implementation setup	129
6.4 The use of patient records for research purposes	139
6.4.1 Accessing the Research Portal Server	140
6.4.2 Releasing patients’ records.....	140
6.4.3 Demonstration Setup	141
6.4.4 Implementation objectives.....	142
6.5 Summary	154
Chapter 7: Discussion	155
7.1 Artefact Evaluation	157
7.1.1 Sharing Information.....	157

7.1.2 Privacy-preservation.....	159
7.2 Research methodology contribution.....	160
7.3 Research Questions Evaluation.....	164
7.3.1 Research Question 1 (RQ1).....	164
7.3.2 Research Question 2 (RQ2).....	169
7.3.3 Research Question 3 (RQ3).....	171
7.4 Summary	173
Chapter 8: Conclusion.....	174
8.1 Thesis Summary	175
8.2 Research Challenges	178
8.3 Research Contribution and Limitations	179
8.4 Directions for Further Research	181
References.....	183
Appendix A: Poster Presented	207
Appendix B: Ethic Application Approval from AUTECH	208
Appendix C: Data Analysis and Coding.....	209
Appendix D: AWS implementation diagram.....	210
Appendix E: Sample of the dummy dataset used	211
Appendix F: ARX Anonymized dataset sample.....	212

List of Figures

Figure (1.1) Thesis Flow	14
Figure (2.1) Anonymization Approaches (Fung, Wang, Chen, & Yu, 2010)	23
Figure (2.2) Privacy threats tree	27
Figure (2.3) Classification of Privacy-Preserving mechanisms in electric health records	38
Figure (3.1) Design Science Research Methodology	44
Figure (3.2) Design Science Research Methodology	45
Figure (3.3) Data Analysis Model (Creswell J. , 2007)	52
Figure (3.4) Creswell’s hierarchical approach (Creswell J. W., 2009)	53
Figure (4.1) Case study findings	74
Figure (5.1) Information sources for the proposed architectural design	76
Figure (5.2) Individual documents that comprise patient information	79
Figure (5.3) documents that contain files	80
Figure (5.4) Information categories and their comprising documents	81
Figure (5.5) Encrypted structured patient’s information	82
Figure (5.6) Proposed Architectural Design	86
Figure (5.7) Information stored on the Requesting Agent for every patient	87
Figure (5.8) Hierarchical Generalization	95
Figure (5.9) Architectural design for sharing healthcare information for research purposes	102
Figure (6.1) Information pre-processing steps performed by the user application	111
Figure (6.2) ROE to store Bob’s information	113
Figure (6.3) Bob’s information stored in the proposed architecture	115
Figure (6.4) Requesting information from CSP and SKA	118
Figure (6.5) Releasing information to the nurse application	119
Figure (6.6) Updating patient information	120
Figure (6.7) Security stations for accessing the system	122
Figure (6.8) Separation of information stored in the system	123
Figure (6.9) Communication channels as a measure of security	124
Figure (6.10) isolated virtual private cloud	129
Figure (6.11) Private and public subnets of the VPC	129
Figure (6.12) Database security measures	130
Figure (6.13) Separation of information stored on the cloud	131
Figure (6.14) Encrypted information that is stored on the cloud	131

Figure (6.15) AWS-KMS used for Trapdoors	131
Figure (6.16) Encrypted document that is stored on the cloud	132
Figure (6.17) Login Page for authentication	132
Figure (6.18) Entering Patient's information	133
Figure (6.19) Script to call the CSR for user authorization	133
Figure (6.20) CSR List of users and their access privileges	134
Figure (6.21) Codes to CSR and SKA	134
(Figure 6.22 a) Indexes of information that is stored on the CSP	135
(Figure 6.22 b) Text Exact Match	135
(Figure 6.23) Unauthorized attempt to down a trapdoor	136
(Figure 6.24) Access denied in response to unauthorized action	136
Figure (6.25) Downloading encrypted document	137
Figure (6.26) Unreadable encrypted document	137
Figure (6.27) Response to unauthorized operation	138
Figure (6.28) Response to unauthenticated user	138
Figure (6.29) Sharing patients records for research purposes	139
Figure (6.30) Releasing patients records for research purposes	144
Figure (7.1) Activity Iterations	155
Figure (7.2) Research Flow in light of DSRM	156
Figure (7.3) Thesis flow in light of the DSRM	162
Figure (7.4) Breakdown of the first research question (RQ1)	165
Figure (7.5) Information that answered RQ1	166
Figure (7.6) Answer breakdown to RQ2	169
Figure (7.7) Derived system characteristics	170
Figure (7.8) RQ3 Answer breakdown	171
Figure (8.1) Dimensional consideration in the proposed architecture	179

List of Tables

Table (2.1) Example of types of attributes in a relational table	28
Table (4.1) Participating Organizations	59
Table (4.2) Interview questions and their relations to main research questions	61
Table (5.1) summary of identified system characteristics	77

Table (5.2) Patients' information categories in the proposed system design	80
Table (5.3) Example of an encrypted index and Trapdoor for encrypted document	83
Table (5.4) Example of users list in stored on the CSR	90
Table (5.5) Sample 2-anonymized dataset	95
Table (5.6) Attribute disclosure using external information	96
Table (5.7) 3-Diverse released patients table	97
Table (5.8) (c, ℓ) -diverse dataset sample	100
Table (5.9) $(2,2)$ -diverse HIV patients block	108
Table (6.1) Description of the dataset used for this instantiation	141
Table (6.2) Sample of the dataset tuples generated in the first stage	146
Table (6.3) Sample of the generated dataset after the insertion of tuples in stage 2	148
Table (6.4) Sample of value-transformed tuples	150
Table (6.5) $(2,2)$ - diverse dataset	152

List of Abbreviations

ABAC	Attribute-Based Access Control
ABE	Attribute-Based Encryption
AES	Advanced Encryption Standard
AMI	Amazon Machine Image
ANSI	American National Standards Institute
API	Application Program Interface
ATC/DDD	Anatomical Therapeutic Chemical Classification Systems with Defined Daily Doses
AUTEC	Auckland University of Technology Ethical Committee
AWS	Amazon Web Services
CAQDAS	Computer-Aided Qualitative Data Analysis Software
CDA	Clinical Document Architecture
CDISD	Clinical Data Interchange Standard Consortium
CDP	Disease Control and Prevention
CEN	European Committee for Standardization
CEN/TC	European Committee for Standardization / Technical Committee
CP-ABE	Cipher Text Policy Attribute-Based Encryption
CSP	Cloud Service Provider
CSR	Cloud Service Registry
DES	Data Encryption Standard
DESE	Deterministic and Efficiently Searchable Encryption
Doc-1	document 1 of patient information
Doc-2	document 2 of patient information
Doc-3	document 3 of patient information
DSRM	Design Science Research Methodology
EC2	Amazon Elastic Compute Cloud
EHII	European Health Information Initiative
ETSI	European Telecommunication Standards Institute
FHIR	Fast Healthcare Interoperability Resources
FISMA	Federal Information Security Management ACT
HER	Electronic Health Record
HIE	Health Information Exchange

HIPPA	Health Insurance Portability and Accountability
HIS	Healthcare Information System
HIV	Human Immunodeficiency Virus
HL7	Health Level Seven
IaaS	Infrastructure as a Service
IAM	Amazon Identity and Access Management
IBAC	Identity-Based Access Control
ICD	International classification diseases
ICT	Information Communication Technology
IEC	international electrotechnical commission
IEEE	Institute of Electrical and Electronics Engineers
IHE	Integrating the Health Enterprise
IHTSDO	International Health Terminology Standards Development Organization
IoT	Internet of Things
IP	Internet Protocol
IS	Information System
ISO	International Organization for Standards
ISO/TC	International Organization for standards/technical Committee
IT	Information Technology
KMS	Amazon Key Management Service
KP-ABE	Key Policy Attribute-Based Encryption
LOINC	Logical Observation Identifiers, Names, and Codes
MAN	Mandatory Access Control
MBAC	Mandatory Based Access Control
NIST	National Institute of Standards and Technology
PaaS	Platform as a Service
PCI DSS	Payment Card Industry Data Security Standards
PEK	Public Key Encryption
PEKS	Public Key Encryption with Keyword Search
PHIN	Publish Health Information Network
PHR	Personal Health Record
PPDP	Privacy-Preserving Data Publishing

PRE	Proxy Re-Encryption
RA	Requesting Agent
RBAC	Role-Based Access Control
RDS	Amazon Relational Database Service
REST	Representational State Transfer
ROE	Request of Enrol
RPS	Research Portal Server
S3	Amazon Simple Storage Service
SaaS	Software as a Service
SDMX-HD	Statistical Data and Metadata Exchange Health Domain
SE	Searchable Encryption
SID	System Identification
Sk	Secret Key
SKA	Secret Key Agent
SKE	Symmetric Key Encryption
Sk_R	Secret Root Key
SLA	Service Level Agreement
SNOMED-C	Systemized Nomenclature of Medicine Clinical Terms
SNOWMED	Systematized Nomenclature of Medicine
SQL	Structured Query Language
SSE	Symmetric Searchable Encryption
UA	User Application
U-IDM	User Identity Management Protocol
USB	Universal Serial Bus
VHIN	Virtual Health Information Network
VPC	Virtual Privacy Cloud
WHO	World Health Organization
WSN	Wireless Sensor Network

Chapter 1: Introduction

With the advancement in information and communications technologies (ICT), it has become easier for healthcare providers to collect and make use of patients' information promptly. These advancements have created new methods to manage patients' information through the digitization of health-related information, they have also contributed significantly towards improving the health care provided to patients at lower costs. Recently, the healthcare sector has shown a growing interest in information technologies. The amount of healthcare records is rapidly growing in detail and diversity and is increasingly collected outside traditional medical record-keeping systems such as within mobile devices, wearable sensors, and home wireless networks (Mamlin & Tierney, 2016). Almost half (48 percent) of healthcare providers polled in a PricewaterhouseCoopers survey said that they had integrated consumer technologies such as wearable health-monitoring devices or operational technologies like automated pharmacy dispensing systems with their IT ecosystems (Compton & Mickelberg, 2014). For instance, The Internet of Things (IoT) and wireless sensor network (WSN) technologies nowadays are considered as a potential solution for healthcare applications. Several researchers focus on designing WSNs for healthcare monitoring systems (Vo, Nghi, Tran, Mai, & Le, 2015).

The IoT is another technology paradigm that is becoming adopted in various applications in the healthcare domain (Islam, Kwak, Kabir, Hossain, & Kwak, 2015). IoT refers to an enormous number of sensors and sensor-enabled devices deployed to collect data about their environment, which frequently includes data related to people. IoT is fundamentally a network of networks with the internet as a backbone. It associates diverse sensors, actuators, and computing systems and communications to provide intelligent services to society (Bandyopadhyay, Balamuralidhar, & Pal, 2013). The automatic exchange of information between two systems or two devices without any manual input is the main objective of the IoT (Borgohain, Kumar, & Sanyal, 2015). The adoption of the IoT concept grants significant help toward collecting and accessing information that was not accessible before in real-time. Areas, which are fast adopting this technology, include industrial monitoring, structural monitoring, environmental monitoring, vehicle telematics, home automation, and healthcare (Rghioui, L'aarje, Elouaai, & Bouhorma, 2014). Healthcare systems are one of the most beneficial applications using wireless medical sensor technologies, which can assist with patient care within homes, work at hospitals, clinics, disaster sites and the open environment (Kumar & Lee, 2012)(Yang, et al., 2014). Several research groups and projects have started to develop

health monitoring systems using WSNs such as CodeBlue (Karla Felix Navarro & Lim, 2009), LiveNet (Sung & Pentland, 2004), CareNet (Jiang, et al., 2008), and Lifeguard (Montgomery, et al., 2004). Such applications generate a massive amount of patients' health-related data leading to a field of big data analytics. The term "Big Data" refers to a large amount of data that traditional database systems cannot process. Big data is a large amount of data that requires new technologies and architectures so that it becomes possible to extract value from it by capturing and analyzing process (Katal, Wazid, & Goudar, 2013). Data from various sensors, hospitals, and social networking sites are a rich source of information for big data (Victor & Lopez, 2016). The healthcare sector has generated massive amounts of data that have huge volumes, enormous velocity, and a vast variety. Such data also comes from various new sources, as hospitals today tend to implement electronic health record (EHR) systems (Patel & Patel, 2016). Big data analytics have started to play a vital role in the evolution of healthcare practices and research. It provides tools to accumulate manage and analyze a huge volume of patients' health-related information produced by healthcare systems (Belle, et al., 2015). Big data analytics in the healthcare domain is currently employed to aid the process of care delivery and disease exploration.

The enhancement of ICT in healthcare is now generating a huge amount of medical data related to several aspects such as diagnosis, testing, monitoring, treatment and health management of patients, billing for healthcare services, and asset-management of healthcare resources (Bock, et al., 2005). eHealth refers to the application of ICT to health, and means of improving health services in terms of access, quality, and efficiency. It is the health-related Internet applications delivering a range of content, connectivity, and clinical care (Maheu, Whitten, & Allen, 2001). eHealth applications are used by doctors, hospitals, insurance providers to record patient health information. These applications are the software and services that manage, transmit, store record information used in healthcare treatment delivery, payments, and record keeping. The eHealth field holds promise to support and enable health behavior change and prevent chronic diseases, it also contributes significantly to improving the healthcare services provided to patients more accurately. For example, wellness data generated by patients using wearable devices or smartphones can be a significant part of a Personal Health Record (PHR). It includes information from the electronic health record (EHR) such as the health conditions of a patient, laboratory results, and medical history. A PHR enables healthcare providers to obtain a much fuller and more reliable record of an individual's health and medical history. It serves as an evolving medical record of treatments provided and their effectiveness as information is added

over time (Etzioni, 2010). The integration of patient-generated wellness data contributes significantly towards a better understanding of patients' health conditions by improving the communication between patients and clinicians (Grossman, Zayas-Cabán, & Kemper, 2009). Several researchers have demonstrated that utilizing patients' wellness data contributes significantly towards healthcare service betterment (Hibbard & Greene, 2013).

1.1 Background

The healthcare industry has generated large amounts of information, driven by record keeping, compliance and regulatory requirements, and (of course) patient care. Information about patients' health generates special value when it is exchanged and collaboratively used among different parties involved in the healthcare area (Kitamura, et al., 2016). Several researchers and interviewed individuals consider immediate access to previously generated medical records during healthcare service delivery as highly important (Fabiana, Ermakovab, & Junghannsa, 2015). Healthcare information systems in healthcare organizations such as hospitals are required to collaborate by exchanging information among medical staff and practitioners for medical care betterment purposes (Gaboury, Bujold, Boon, & Moher, 2009). The definition of the term "collaboration" in the field of healthcare includes the concept of sensibly sharing a collective perspective that includes information, norms, social expectation, activity goals, and meaning. It is the communication that occurs among healthcare practitioners when sharing information and skills regarding patient care (Weir, et al., 2011).

In the healthcare domain, patients usually acquire medical care from a wide range of caregivers based on their proximity, quality of care received, cultural attitudes, and bedside manner. Medical care may be received from various caregivers such as hospitals, pharmacy, laboratory, physician groups, nurses, school clinics, and public health places (Thompson & Brailer, 2004). This has led to the fragmentation of patients' information in heterogeneous systems. The majority of this collected information is stored in heterogeneous distributed health information systems which are mainly proprietary (Kokkinaki, Chouvarda, & Maglaveras, 2006), and as a consequence, health-related information stored in these systems cannot be easily accessed to present a clear and complete picture of an individual patient when needed. For example, when a patient visits a healthcare provider such as a general practitioner, he or she often requires additional medical services or attention over some time whether it is specialized medical examination such as magnetic resonance imaging scans, or a routine medical examination such as cholesterol test and blood sugar checks.

A survey conducted by Software Advice found that 46 percent of patients want their doctors to directly exchange their health-related records while 21 percent preferred in-person delivery. When patients were asked about the way their medical records were shared among multiple healthcare providers, only 39 percent of patients said providers directly exchange records, and 25 percent had to deliver a physical document to other healthcare providers themselves. Such finding illustrates the challenge patients faced when they shared or obtain their medical records while using multiple healthcare providers (Pennic, 2015). A study in an outpatient clinic found that pertinent patient data were unavailable in 81% of cases; the entire medical record was unavailable 5% of the time with an average of four missing items per case (Walker, et al., 2004); these findings point to a need for having a certain mechanism to enable the sharing of patients' health information, and to achieve efficient collaboration among entities involved in the healthcare domain. The extensive information exchange in the healthcare domain takes place among primary and secondary healthcare providers in two flow directions as described in (Casola, Castiglione, Choo, & Esposito, 2016). The first communication flow takes place when secondary healthcare providers retrieve data about patients to provide the appropriate follow-up examination such as specialist medical services and examinations, while the second communication flow happens when primary healthcare providers are notified whenever new information such as medical records relating to a patient becomes available. Another flow of information that takes place at the administration level, for example, collecting relevant information for a range of administration-related functions such as billing. The concept of sharing information in the healthcare domain helps to better understand the health needs and therefore improve the quality of care provided to patients (Kitamura, et al., 2016). For that, the seamless exchange of multimedia clinical information is considered as a fundamental requirement. Different technological approaches can be adopted for enabling the communication and sharing of health records segments (Tsiknakis, Katehakis, & Orphanoudakis, 2002).

Nevertheless, the use of collected data is a valuable source for analysis that benefits both medical research and practice. It leads to effective ways of preventing and managing illnesses, as well as the discovery of new drugs and therapies, however, many challenges need to be overcome before obtaining the best of what sharing information in the healthcare can offer. For example, sharing healthcare information across different parties in healthcare increases concerns related to security, privacy, integrity, and confidentiality of healthcare data. The information in the healthcare domain may contain commonly considered private information

that may concern patients when sharing it with other parties. Patients require their information to remain always secure and private as a condition for granting permission to share it among different parties. In (Whiddett, Hunter, Engelbrecht, & Handy, 2006), widespread patient consultation, including NZ patients, found high levels of support for sharing their health-related information provided that such information remains secure. Personal information refers to the information that includes factual or subjective information about an identifiable individual. Information privacy refers to an aspect of information technology that deals with the ability that an organization or individual has to determine what data in a computer system can be shared with third parties. It is the flow of information according to social norms, as governed by context (Nissenbaum, 2009). The privacy of information exists when the usage, release, and circulation of personal information are controlled (Culnan, 1993). Several privacy-related laws and policies are enforced in almost every social setting to preserve the privacy of individuals' information. The share of healthcare information conflicts with two main ethical issues, which are privacy and security (Denecke, et al., 2015). In (Deering, 2013), the author briefly outlined several concerns that may arise among both health care providers and patients due to receiving data from patients about their health outside the clinical visit. The author also outlines several technical issues related to the capture, transmission, and integration of the data. For example, standardization is a challenge that is currently hindering the integration of data from the various health application systems. Information should not only be received but also understood. In fact, the utility of the current advancement of ICT in the healthcare domain is still in the early stages, many challenges require overcoming before obtaining the best of what such advancements can offer.

1.1.1 Interoperability Challenge

Interoperability is the ability to share and use information across multiple system technologies seamlessly. Interoperability is a fundamental requirement for the health care system to derive the societal benefits promised by the adoption of electronic healthcare records (Brailer, 2005). The seamless exchange of vital information among healthcare practitioners played a significant role in reducing medical errors and facilitated better integration of health-related records (Iroju, Soriyan, Gambo, & Olaleke, 2013). To realize the full potential of collected medical data, health-related IT systems and products are required to share information seamlessly among each other, but unfortunately, the vast majority of medical devices, electronic health records, and other IT systems lack interoperability. The authors in (Whitman & Panetto, 2006) defined four levels of interoperability namely; technical, syntactic, semantic, and organizational; a

similar definition was given by the European Telecommunication Standards Institute (ETSI) (Veer & Wiles, 2008). Technical interoperability refers to the ability of heterogeneous systems to exchange data without guaranteeing the ability of the receiving system to understand the data in a meaningful way. Syntactic interoperability is the preservation of the clinical purpose of the data during transmission among healthcare systems. Semantic interoperability refers to the ability of systems to interpret the information that has been exchanged similarly through a pre-defined shared meaning of concepts while organizational interoperability refers to the ability to facilitate the integration of business processes and workflows beyond the boundaries of a single organization. In (Diaz, 2016), the author states that sharing data in a useful way in the healthcare domain is impossible without semantic interoperability among disparate healthcare IT systems. Semantic interoperability deals with the content of the message exchanged among health information technology systems. It is about the ability of systems to understand the meaning of the shared data. Interoperability is important because treatment and health care providers have increased and become more specialized, and patients have become mobile. Such large-scale adoption of electronic healthcare applications requires semantic interoperability (Sachdeva & Bhalla, 2010).

Patients' health records are often stored in a non-standard, non-coded, structured and non-structured form hindering the exchange of information among health information systems (Lau & Shakib, 2005). It is currently a major challenge in the healthcare industry to achieve interoperability among proprietary applications provided by different vendors (Cantwell & McDermott, 2016). For instance, a hospital may use one or more applications to share clinical and administrative information, and each application may support multiple communication interfaces and protocols that must be modified and maintained. Adopting common data structures within the healthcare organizations is a decision that has been met with reluctance due to financial concerns and other barriers related to changing the existing workflow and staff training costs (Gabriel, Furukawa, Jones, King, & Samy, 2014). One of the primary reasons for this reluctance is the inability of the electronic health records to interlink and communicate with each other due to the lack of comprehensive data standard that facilitates the exchange of data using a common data model (Bowles, et al., 2013). The inability of healthcare information systems (HISs) to interoperate on the national scale reaps the full benefits of e-health (ITU, 2011). In (Iroju, Soriyan, Gambo, & Olaleke, 2013), the authors aimed to upraise the concepts of interoperability in the context of healthcare, its benefits, and challenges. The authors write: "However, as beneficial as data interoperability is to healthcare, at present, it is largely an

unreached goal”. This is primarily because electronic healthcare information systems used within healthcare organizations have been developed independently with diverse and heterogeneous ICT tools, methods, processes, and procedures. This leads to generating a large number of heterogeneous and distributed proprietary models for representing and recording patients’ information.

Heterogeneity is considered a major obstacle to healthcare information systems’ interoperability. Healthcare information systems differ from application to another and from a country to another. This means that the structure of healthcare records and the methods used for exchanging their contents may significantly vary. Due to the existence of various independent data standards repositories such as LOINC (Logical Observation Identifiers, Names, and Codes), ICD (International classification diseases), and SNOWMED (Systematized Nomenclature of Medicine), it is not possible for healthcare facilities to successfully achieve interoperability. There is no unified standardization format that can act as a single comprehensive standard for data interpretation and translation of medical vocabulary and terminologies (Ogunyemi, Meeker, Kim, & Boxwala, 2013). For that, the solution is expected in the standardization of electronic health information structure, content, and the way of exchanging them (Gross, 2005). Currently, it is determined that there is no existing single data standardization structure that can effectively share and interpret patient data within heterogeneous systems (Blackman, 2017).

Utilizing information from various systems and environments in the healthcare industry adds significant value to the field of healthcare. Information is today collected from different sources and heterogonous systems, which require aggregation to make use of it. For this aggregation to happen, it is important to make sure that patients have permitted to share their health-related information. Information privacy is a key reason behind the patients’ rejection to share their health information. On the other hand, the aggregation of information from various systems requires transmitting information from a source to another however, due to the lack of standardization and therefore poor interoperability, it becomes not possible to automatically transmit information from a system to another. Privacy and interoperability can drive the healthcare sector to a better position in terms of information utility.

1.1.2 Privacy Challenge

The continuous advances in information technology have reduced the amount of control over personal data and opened up the possibility of a range of negative consequences as a result of

access to personal information (Van-den Hoven, et al., 2016). Privacy-related regulations are considered one of the biggest challenges to health data sharing; they prohibit the transmission and distribution of personal health information even among collaborating organizations impeding research and reducing the utility of the datasets (Ezea & Peyton, 2015). Due to privacy concerns and the lack of healthcare information sharing as a consequence of it, most of the facilities aim at building clinical decision support systems using a limited amount of patient data from their healthcare information systems to provide important diagnosis relation decisions. Moreover, it becomes infeasible for a newly established healthcare facility to build a robust decision-making system due to the lack of sufficient patient records required to train such decision-making models (Li, Bai, & Reddy, 2016). According to the Privacy Act 1993 (New Zealand), personal information should be collected directly from the individual, unless they have authorized another person to pass on their information, or if it is not reasonably practical in the circumstances.

Healthcare systems contain sensitive information that must be managed in a privacy-preserving way. For that, it is a mandatory step to adhere to legal frameworks such as the Health Insurance Portability and Accountability Act (HIPPA) (Public Law, 1996) and the Data Protection Act (Gunasekara & Dillon, 2008). Such frameworks specify the responsibilities of organizations with regards to the privacy protection of personal health information. However, complying with these frameworks is both challenging and costly for healthcare organizations (Gkoulalas-Divanis & Loukides, 2015). Several attempts have been made by researchers to allow the exchange of medical information among medical practitioners/data analysts in a privacy-preserving manner. The main privacy challenge remains in the management of this collected data which is still largely unaddressed (Weber, 2015). There are many policy-related issues such as privacy policies that must be addressed to realize the full potential of sharing healthcare information (Hripesak, et al., 2014) (Gkoulalas-Divanis & Loukides, 2015). In (Rashid & Yasin, 2015), the authors state that sharing healthcare information using healthcare information systems based on privacy preservation rarely handles healthcare information sharing among healthcare-related entities at different places; therefore, there is a need to address such collaboration based on privacy preservation.

1.2 Research Problem

Due to the diversity and complexity of the existing healthcare structure, in which patients' health information is distributed to multiple entities such as hospitals, healthcare centres, and

cloud servers, an appropriate architecture is one of the most important design issues for sharing healthcare information in a privacy-preserving manner.

A centralized architecture design would not be convenient due to the lack of interoperability of the vast majority of healthcare information systems. Interoperability is defined as the ability to share and use information across multiple systems seamlessly (Oude, Velsen, Huygens, & Hermens, 2015). Currently, it is determined that there is no existing single data standardization structure that can effectively share and interpret patient data within heterogeneous systems (Blackman, 2017).

Despite the use of information technology solutions in the healthcare industry, there are various challenges encountered such as the high infrastructure management costs, dynamic needs for computational resources, scalability multi-tenancy, and increased demand for collaboration (Priyanga.P & MuthuKumar.V.P, 2015). The advancement in the healthcare industry requires modernizing healthcare information systems to facilitate collaboration and coordination among parties involved in the healthcare domain at lower costs. In healthcare, the availability of information regardless of the location of the patient and the clinician is a key driver towards patients' satisfaction and healthcare service betterment. For that, there is a stressing need for having a decentralized design of the architecture for healthcare information systems that allows for asynchronous interactions among parties involved in the healthcare domain concerning privacy regulation (Casola, Castiglione, Choo, & Esposito, 2016).

Cloud Computing

Cloud computing appears to be the dreamed vision of the healthcare industry; it matches the need of healthcare information sharing directly to various healthcare-related parties over the internet, regardless of their location and the amount of data being shared (Guo, Kuo, & Sahama, 2012). Health information exchanges enable healthcare organizations to share data contained in largely proprietary information systems. Cloud computing technology is seen as a potential solution for enabling healthcare organizations to focus their efforts on clinically relevant services and improved patient outcomes (Kuo, 2011). Cloud Computing is an emerging new computing paradigm designed to deliver computing resources and services through networked media such as the Web (Sultan, 2014). It is a computing paradigm in which resources of the computing infrastructure are provided as a service over the internet (Yu, Wang, Ren, & Lou, 2010). In the simplest terms, cloud computing refers to means of storing and accessing data and programs over the internet instead of the computer's hard drive (Griffith, 2016).

The technology of cloud computing enables relatively new business models in the computing world. It offers functionality for managing information data in a distributed, ubiquitous, and on-demand network access to a shared pool of configurable computing resources (Mell & Grance, 2011). Resources in cloud computing can be rapidly provisioned and released with minimal management effort supporting several platforms, systems, and applications (Doukas, Pliakas, & Maglogiannis, 2010). Cloud computing is an attractive paradigm of computing for the healthcare domain, due to the elasticity of resources and reduction of the operational costs. This allows for new ways of developing, delivering, and using healthcare services (Griebel, et al., 2015). Cloud computing offers practical solutions in the healthcare domain and sharing information is one of them (Zhang & Liu, 2010). For example, the Collaboration Care Solution is a system developed by IBM and Active Health Management in 2010. The cloud-based system enabled medical and healthcare staff to easily access healthcare data and information from different sources. The system was beneficial for patients, who were suffering from chronic conditions, to connect with their physicians, and follow up their prescribed medications and treatment (Aziz & Guled, 2016).

However, despite the advantages that cloud computing offers to the healthcare domain; privacy protection is a major challenge (Yüksel, Küpçü, & Özkasap, 2017). Such concerns are caused by the fact of having medical data and information that is classified as confidential, stored in cloud servers, a virtual world where information can be easily hacked (Aziz & Guled, 2016). From the consumers' perspective, privacy when storing and sharing health-related information on the cloud is a primary concern, because data is stored in different places. Such concern prohibits the adoption of cloud computing in the healthcare domain (Chen & Zhao, 2012) (Shariati, Abouzarjomehri, & Ahmadzadegan, 2015).

Information privacy is the desire of individuals to control or have some influence over data about themselves (Bélanger & Crossler, 2011). It is, in other words, the right of individuals to determine how and to what extent information they communicate to others is used. Healthcare data includes sensitive records that should not be made available to unauthorized people to protect the privacy of patients. Information privacy protection is very essential to build users' trust in order to reach the full potential of cloud computing in the healthcare domain. For that, an important characteristic in healthcare cloud-based information systems is the ability to assure patients that their data is protected in the cloud, and their private information will only be disclosed to responsible parties.

1.3 Research Motivation and Significance

The technology of cloud computing appears to be the dreamed vision of the healthcare domain in terms of sharing information and collaboratively using it for healthcare services and research. Cloud computing is an attractive diagram of computing that enables new ways of delivering healthcare services. However, despite the advantages that cloud computing offers to the healthcare domain; privacy protection is a major challenge (Yüksel, Küpçü, & Özkasap, 2017). Such concern arises due to storing information that is considered highly sensitive in the virtual world (cloud servers) where it can be easily hacked or accessed by unauthorized persons (Aziz & Guled, 2016). Therefore, privacy when storing and sharing health-related information on the cloud is a primary concern, because data is stored in different places. Such concern prohibits the adoption of cloud computing in the healthcare domain (Chen & Zhao, 2012) (Shariati, Abouzarjomehri, & Ahmadzadegan, 2015).

Such limitation has motivated the researcher to review the current literature in the body of knowledge in relation to cloud computing, identity and access management, encryption schemes, and privacy issues in the cloud. The motivation of the researcher was about finding a way to adopt cloud computing technology in healthcare information systems without violating the privacy of information. The ultimate goal was to address the privacy concern that arises due to storing sensitive information in a cloud environment concerning legal frameworks such as HIIPA (Public Law, 1996) and Data Protection Act (Gunasekara & Dillon, 2008).

This research intended to contribute to the overall knowledge about how the technology of cloud computing can be adopted by healthcare information systems. The outcome of this research was expected to allow the adoption of cloud computing in the healthcare domain. This would enable collaborative and privacy-preserving use of patients' information to improve the services provided to patients. Since an appropriate architecture is one of the most important design issues for sharing healthcare information in a privacy-preserving manner, the intention of this research was to design a decentralized cloud-architecture for healthcare information systems that allows for asynchronous interactions among parties involved in the healthcare domain with respect to the privacy regulations.

1.4 Research Questions

Healthcare information systems play a vital role in the quality of care provided to patients; however, the utility of such systems in terms of sharing information is hindered and considered

a bleeding-edge in the information technology field. Privacy is a major challenge towards gaining the trust of patients when sharing their records among responsible parties. To gain patients' trust and acceptance to share their health-related information, there is a stressing need to design privacy mechanisms that enable the share of healthcare information in a privacy-preserving manner. The main intention of this research is to design a cloud-based architecture for healthcare information systems, to facilitate collaborative use of patient information among the various parties involved in the healthcare domain in a privacy-preserving manner. The scope of the research aimed to answer the following questions:

1. How do we maintain the privacy requirements of healthcare data while it is stored on the cloud?
2. What are the characteristics of a privacy-preserving cloud-based architecture for sharing healthcare information?
3. What information can be disclosed for statistical analysis by cloud providers?

1.5 Thesis Publications

To validate the contribution of this thesis to the body of the knowledge, the work conducted in this thesis has been peer-reviewed through publishing a book chapter, conference article, and presenting a poster at a conference.

Book Chapter: The book chapter aimed to discuss the main challenges encountered before healthcare information systems can collaboratively share patients' records. The chapter was a result of an intensive literature review in the area of collaborative use of patients' information using the current information systems. The chapter was concluded by identifying gaps in the literature and outlining potential research directions for enabling the share of information in the healthcare domain in a privacy-preserving manner.

- Alhaddadin, F., Gutiérrez, J. A., & Liu, W. (2018). The collaborative use of patients' health-related information: Challenges and research problems in a networked world. In *D. Saha (Ed.) Advance in Data Communications and Networking for Digital Business Transformation* (pp. 227-271). IGI Global.

Conference Paper: The conference article aimed to explain the proposed architectural design in terms of its fundamental aspects and components. The article aimed to elaborate on how the proposed cloud architectural design overcomes the challenges encountered when adopting cloud computing technology in the healthcare domain.

- Alhaddadin, F. (2019). Privacy-aware cloud-based architecture for sharing healthcare information. *International Conference on Information Resources Management (CONF-IRM)*. Auckland, New Zealand

Poster: The poster was presented in the International Conference on Information Resources Management (CONF-IRM) in 2019 at the Auckland University of Technology. The poster included a breakdown of the aspects involved in the proposed architectural design, and further explained how the integration of these aspects enables for collaborative use of patients' information in a privacy-preserving manner. The different aspects of the proposed design were reviewed by a number of academics and researchers in the areas of cloud computing and healthcare information systems. The poster is annexed at the end of this thesis in Appendix A.

1.6 Thesis Structure

This thesis is presented in eight chapters. The **first chapter** presented an overview of the research that includes: the current situation of healthcare information systems in terms of their ability to share information, barriers that hinder the collaborative use of healthcare information, the research problems and questions that framed this research, and finally the significance of this research. These provide an introduction and overview of the research. The **following chapter** presents a theoretical review of the literature. It concentrates on the efforts that have been put by researchers to enable the current information systems to collaboratively share and use information in the healthcare domain. Privacy, interoperability, identity and access management, and encryption are the main topics of the chapter. The chapter also reviewed the efforts that researchers have put to enable the adoption of cloud computing in the healthcare domain.

Chapter three presents the design and the methodology followed for this research. The chapter discusses the problem area and how the objectives of solutions are identified. The research problem has been observed and suggested as future research opportunities in the literature. The case study approach was followed for the conduction of this research which is also explained in the chapter.

Chapter four presents the process of conducting the case study research activities that include gathering data from research participants, organizing, and analyzing it. It also presents the discussion of the data analysis findings in terms of identifying the objectives of the solution.

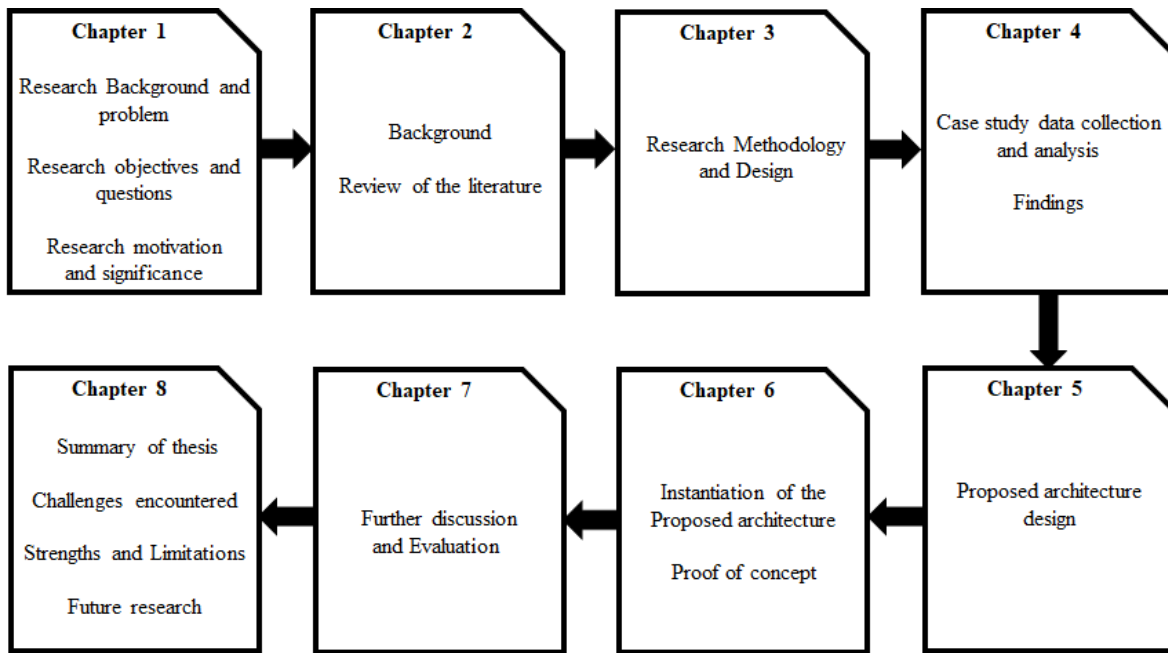


Figure (1.1) Thesis Flow

The proposed solution which is cloud architecture for the healthcare domain is presented in **Chapter five**. The chapter presents the proposed cloud architecture in terms of its components and key aspects of its design. The chapter presents an explanation of how each component works in the proposed design to achieve storing and sharing healthcare information for medical treatment purposes and research purposes in a privacy-preserving manner.

Chapter six presents a demonstration of the proposed architectural design. The chapter presents a scenario-based instantiation of the proposed architectural design to demonstrate its ability to store and share information in a privacy-preserving manner. The implementation of the system is also presented and explained as a proof of concept to the proposed cloud architecture. **Chapter seven** presents a discussion of the research findings. The chapter discusses three main aspects of the research which are: the contribution of the research methodology towards the success of this research, evaluation of how research questions were answered, and finally evaluation of the designed cloud architecture in terms of its ability to share healthcare information in a privacy-preserving manner.

Finally, **Chapter eight** presents the conclusion of the thesis. The chapter presents a summary of the research, identified research challenges, and limitations of the research. Areas for future work arising is also listed in the chapter. A full list of references is presented, and finally, a list of support documents is provided in the Appendix at the end of the thesis

Chapter 2: Literature Review

With the growing use of information and communication technology in the healthcare sector, the issue of accessing and sharing information is becoming increasingly important. Among all shared information, healthcare information has received considerable attention from researchers and individual users (Torabi & Beznosov, 2013). Nowadays, exchange and share of clinical information among Information Systems (IS) are becoming one of the main ways to improve the quality of the services provided to patients (Peixoto, Domingues, & Fernandes, 2016).

Governments' policies and actions nowadays support the adoption of health information exchange (HIE) for the goal of improving the healthcare services provided to patients by addressing fragmented personal health information. At present, there is a mix of both paper and electronic medical records in use, which may be held by multiple healthcare entities. This results in a fragmented picture of an individuals' health history and an incomplete potential to adversely impact clinical decisions limiting opportunities for proactive healthcare, such as prevention and healthcare promotion activities across multiple agencies (Vest & Gamm, 2010). The inability of healthcare services to quickly and easily access patient health information can compromise treatment and care decision-making especially when urgent treatment is required (Naylor, 2010).

Many efforts have been put towards facilitating the share of information in the healthcare sector around the world. Countries around the world are continuously investing in health information and communications technologies (ICTs) as critical tools for improving their healthcare services (Adler-Milstein, Sarma, Woskie, & Jha, 2014). For example, in the united states, the Public Health Information Network (PHIN) is an initiative developed by the Centre for Disease Control and Prevention (CDC) to establish and implement a framework for sharing public health information electronically (Rouse, Margaret, 2010). The main goal of the network is threefold: (1) to facilitate communication among public health practitioners throughout the United States, (2) to make information accessible, and (3) to make secure data exchange as swift and smooth as contemporary technology will allow (Baker, Friede, Moulton, & Ross, 1995). The Virtual Health Information Network (VHIN) for New Zealand is another attempt that aims to create and sustain an environment that captures value from linking health data collections, through world-leading health research, policy development and service planning (Olds, 2015). The VHIN project aims to build capacity and capability, create easily accessible

resources, support open sharing of code and resources, contribute to improved data quality, and undertake high-quality research in the healthcare sector. The European Health Information Initiative (EHII) is also another example. It is a World Health Organization (WHO) network committee for improving the information that underpins health policies in the European Region. The EHII network aims to foster international cooperation to support the exchange of expertise, build capacity, and harmonize processes in data collection and reporting (World Health Organization , 2017).

There is a large number of other ICT networks and projects in the world that aim to facilitate the exchange of information in the healthcare domain to improve healthcare services provided to patients, however, their efficiency in sharing healthcare information is limited due to several challenges. Interoperability and anonymization are considered two major challenges that need to be addressed in order to gain the benefits of sharing healthcare information. Interoperability is the ability for two or more systems or components to exchange information and use the information that has been exchanged (Oude, Velsen, Huygens, & Hermens, 2015), while anonymization is defined as the process by which personal information is altered in a way that an individual patient can no longer be identified directly or indirectly (Victor & Lopez, 2016). Anonymization is important when using aggregated patients' healthcare information for research purposes to protect the privacy of patients' health information.

2.1 Interoperability

Interoperability is defined as the ability for two or more systems or components to exchange information and use the information that has been exchanged (Oude, Velsen, Huygens, & Hermens, 2015). Often, patients' health records are stored in a non-coded, non-standard, structured, and non-structured form, and hinders the exchange of information among health information systems (Lau & Shakib, 2005). Medical institutions comprise a large variety of operational systems supplied by different vendors which include core systems such as electronic medical records, order entry and medical accounting systems as well as departmental systems such as clinical examinations, radiation information management, and medical image management systems. The information handled across these systems extends very widely (Natsuki, 2008). Achieving interoperability among proprietary applications provided by different vendors is currently considered as a major challenge in the healthcare industry (Cantwell & McDermott, 2016).

To achieve interoperability among different systems in the healthcare domain, several efforts have been put by various desperate parties. Medical information standardization has been considered as a solution to interoperability across healthcare information systems (Natsuki, 2008). The International Organization for Standardization (ISO) defined standard as a document, established by consensus and approved by a recognized body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context (ISO, 2004). The concept of standardization grants a number of benefits including preventing from single vendor lock-in, promoting a healthy market competition with associate cost savings, reducing the risks of new technology development and removing the need for expensive customized solutions (Meingast, Roosta, & Sastry, 2006) (Wager, Lee, & Glaser, 2013). Standardization is an important aspect of enabling the use of networks to share and utilize medical and healthcare information. Healthcare-related information usually comes in various forms; not only names of diseases, drugs, and treatments but also other data forms such as images, numerical values of examination results, graphs, and text. Therefore, assuring the connection between different systems among institutions, regions, and nations requires integration to details such as terminology, encoding, protocol, and security (Ishigure, 2017).

Standards in general fall into two broad groups; proprietary standards and open standards (Adebesin, Foster, Kotz, & Greunen, 2013). Proprietary standards are developed for private use by profit-driven industry organizations. The specifications of such standards remain unrevealed and are subject to copyright law. Open standards are open for use by all interested stakeholders. They can be developed by for-profit and non-profit organizations. The standard specifications and necessary documentations are made available for public use, either free of charge or at a nominal fee.

There are various efforts that organizations globally have made to develop interoperability standards for healthcare systems. For example, the International Organization for Standardization (IOS) which is the world's largest developer of the standard has developed 162 national standards bodies globally (ISO, 2017). The standards of ISO are developed by various group members in different technical committees that are made up of national member bodies. Memberships offered by ISO are categorized into three main categories namely, full, correspondent, and subscriber (ISO, 2017). Each membership category has a different accessibility degree to ISO's standards, participation, and development. E-health Standards are developed by ISO's health informatics technical committee, ISO/TC 215. The standards are

meant to support the growth in the use of information and communication technology in the healthcare domain to facilitate the secure and seamless exchange of health-related information that is accessible to authorized users when required (ISO, 2013).

The World Health Organization (WHO) publishes and maintains the codes of International Classification of Diseases (ICD) for classification of diseases, health conditions and causes of death (WHO, 2017), the Anatomical Therapeutic Chemical Classification Systems with Defined Daily Doses (ATC/DDD) provides codes for the classification of medicines (WHO, 2017), and the Statistical Data and Metadata Exchange Health Domain (SDMX-HD), a standard for the exchange of health indicators (SDMX-HD, 2016) among others. World Health Organization also collaborates with the International Health Terminology Standards Development Organization (IHTSDO) to enable cross-mapping of the Systemized Nomenclature of Medicine Clinical Terms (SNOMED-CT) terminologies with ICD codes (WHO, 2017).

The European Committee for Standardization (CEN) is another non-profit organization that aims to develop standards for the goal of removing trade barriers across European countries through coordination of the development of European standards (CEN, 2012). CEN comprises national standard bodies of 27 European Union countries in which these standards are adopted as national standards. There is also an agreement of cooperation between ISO and CEN that aims to prevent the development of conflicting or parallel standards. In this agreement, the standards of ISO can be adopted as CEN standards and vice versa. The e-health standards of CEN are developed by the health informatics technical committee, CEN/TC 251 (CEN, 2009). The goal is to facilitate the adoption of standards that can potentially enable organizations in Europe to optimally use their health informatics systems, via the development and adoption of international standards. The CEN/TC 251 also collaborates with other standards development organizations such as ISO/TC 215, the Clinical Data Interchange Standard Consortium (CDISC), Health Level Seven (HL7), and the IHTSDO. The Clinical Data Interchange Standard Consortium (CDISC) is a non-profit organization that is open, and multidisciplinary (CDISC, 2013). The major goal of CDISC is to develop standards to support the acquisition, exchange, submission, and archive clinical research data and metadata. CDISC aims at developing platform-independent standards that facilitate the interoperability of information systems to improve research in the healthcare field. CDISC collaborates with HL7 via an agreement with the latter to facilitate the harmonization of their clinical research standards (CDISC, 2013). Health Level Seven (HL7) is an American non-profit organization accredited

by the American National Standards Institute (ANSI) that develops standards for exchanging clinical and administrative data among heterogeneous healthcare applications (HL7, 2017). HL7 has a variety of membership categories such as individual, organizational, caregiver, students, and supporter. Each membership category offers a range of different benefits. The standards of HL7 are developed by volunteers who work in various working groups, under the stewardship of the technical steering committee (Benson, 2012). HL7 also collaborates with other standard developing organizations including CEN, ASTM International, ISO, and IHTSDO (HL7, 2017). The Institute of Electrical and Electronics Engineers (IEEE) is known as the largest professional association in the world that aims to advance technological innovation and excellence for the benefits of humanity (IEEE, 2017). IEEE also as part of its work develops standards for a range of products and services. It also develops standards for IT healthcare devices to facilitate the interoperability of medical devices (IEEE, 2017). IEEE also cooperates with other standards developing organizations such as ISO, the international electrotechnical commission (IEC), on the joint development of international standards (IEEE, 2017).

Many other organizations are involved in the development of interoperability standards in the healthcare domain such as National Electrical Manufacturers (DICOM, 2011), ASTM International (ASTM, 2012), Integrating the Health Enterprise (IHE) (IHE, 2016). Each of these organizations focuses on developing standards for healthcare information exchange for the goal of achieving interoperability between healthcare information systems.

FHIR Standard is another standard proposed by Health Level Seven (HL7) as a response to the issue of interoperability in healthcare information systems. Health Level Seven (HL7) has provided a series of frameworks for the exchange, integration, and search of medical health information and has developed standards to resolve interoperability between systems. CDA, V2 Message, and V3 Rim are of the main standards developed and proposed by HL7 (Begoyan, 2007). CDA (Clinical Document Architecture) is an XML-based mark-up standard intended to specify the encoding, structure, and semantics of clinical documents for exchange (Rouse, 2015). The HL7 V2 standard was firstly developed in the early 1990s, and it is widely used nowadays. It is a messaging standard that allows the exchange of clinical data between systems. It was designed to support a central patient care system as well as a more distributed environment where data resides in departmental systems. However, the drawback of the HL7 V2 standard is that it takes a long time to develop various services based on HL7 V2. It also lacks an information transfer that ensures semantic interoperability. Therefore, applications

participating in communication using HL7 V2 must have mutual agreements to achieve interoperability (Begoyan, 2007).

HL7 V3 Rim is another standard that was developed in 2005 to overcoming the drawbacks of the previous version (V2). It ensured interoperability and used XML technology and object-oriented approaches. However, the development using this standard was not easy due to the complexity of medical information and difficulties in modelling the complete services of engineers without professional knowledge (West, 2015).

To overcome the drawbacks of the previously mentioned versions, HL7 introduced Fast Healthcare Interoperability Resources (FHIR) as the next generational standard for sharing healthcare records. It is a new standard framework that is based on previous data format standards and utilizes the beneficial elements of HL7-Version 2 and HL7-Version 3 (HL7, 2016). FHIR is a standard that is based on Representational State Transfer (REST) architecture style that enables it to be extended to mobile and other light-weight devices. As a result, the interface can provide services that can be accessible to various healthcare-related practitioners such as pharmacists, doctors, and patients (HL7, 2015). The authors in (Lee, Kim, & Lee) list a number of improved functions in the FHIR compared with the existing standards which include a strong focus on implementation, multiple implementation libraries, specification is free to use with no restrictions, interoperability out-of-the-box base resources can be used as is with adaptability for local requirements, evolutionary development path from HL7 Version 2 and CDA standards, a strong foundation in web standards such as XML, JSON, HTTP and OAuth, RESTful architectures support, seamless exchange of information using messages or documents, concise and easily understood specifications, human-readable wire format for ease of use, and finally solid ontology-based analysis with rigorous formal mapping for correctness. More information about FHIR's improved functions can be found in (HL7, 2016).

FHIR is today gaining widespread attention for its potential to foster innovative approaches to sharing clinical data using very modern web technology-based ideas. It is attractive due to its relatively easy implementation; it comprises a set of modular components called resources that can easily and incrementally be assembled into working systems (Alterovitz & Yao, 2015). In (Ahier, 2015), the author writes "FHIR is not simply adding additional standards to an already overflowing kettle, but rather the next step in the evolution of standards that will truly promote interoperability.". In (HIMSS, 2016), Russel Leftwich who serves on the HL7 board believes FHIR-based applications will spread rapidly as the standard matures. He likens the standard's

maturity journey to the evolution of the iPhone, where capabilities and use will increase with each successive version. “The potential for what it will be able to support over the next few years is tremendous,” said Leftwich, a senior clinical advisor for interoperability at InterSystems, and serves as an adjunct assistant professor of Biomedical Informatics and Vanderbilt University School of Medicine.

Several efforts have been put by researchers for the goal of adopting FHIR standards integration in various healthcare-related information systems. In (Alterovitz, et al., 2015), the researchers aimed to link genome and phenome variants to patient’s electronic health records to eventually support clinical decision support systems. The main intention of the research was to unify how genomic variant data are accessed from different sequencing systems. The scope of the research aimed to develop a specification for the basis of a clinic-genomic standard that builds upon FHIR. The research resulted in a successful design, deployment, and use of the Application Programming Interface and was demonstrated and adopted by the HL7 Clinical Genomics Workgroup. The feasibility was demonstrated by developing three apps by various types of users with background levels and locations. The research concluded that an entire data (and web) standards-based approach could prove both effectiveness and efficiency for advancing personalized medicine. In (Khalilia, Choi, Henderson, Iyengar, Braunstein, & Sun, 2015), the authors demonstrated a software architecture for developing and deploying clinical predictive models using web services via FHIR standard. The resulting predictive models were deployed as FHIR resources that receive requests of patient information, perform prediction against the deployed model, and respond with prediction scores. The response and prediction time of the FHIR modelling web services were evaluated to assess the practicality of the approach. The research found that the system was reasonably fast with one second total response time per patient prediction. Another research conducted in (Franz, Schuler, & Krauss, 2015) aimed to show an integrated monitoring solution based on Continua and Integrating the Healthcare Enterprise, which was tested by more than 130 patients and 14 healthcare institutions. The low battery life of smartphones due to high data traffic was the trigger to conduct the research. The research found that there was a significant decrease in data traffic when relying on a RESTful architecture in combination with FHIR, due to the efficient resource handling of web service connections that FHIR offers.

However, the interoperability of electronic information remains a tremendous challenge especially with over 100 electronic healthcare information standards that currently exist and used (Ogunyemi, Meeker, Kim, & Boxwala, 2013). As the need to exchange healthcare

information continues to grow rapidly, the sharing and communicating health-related information across healthcare information systems becomes impossible due to the variety of data standardization models employed by the healthcare information systems which can only ensure interoperability within its open operational domain. Currently, there is no single source data standardization model to achieve semantic health data interoperability between heterogeneous systems (Sinaci & Erturkmen, 2013)(Blackman, 2017). In (Khan, et al., 2014), the authors write “Data interoperability is also impossible to accomplish in the current state due to the lack of a relationship between healthcare data and the different health information systems, a growing concern for healthcare practitioners and facilities since it prevents the provision of better patient care”. Currently, there is no existing model that is implemented to support the different vocabularies, data interpretation algorithms, and mapping tools in a single source environment; they are all stand-alone applications that hinder interoperability among heterogeneous systems (Sinaci & Erturkmen, 2013).

2.2 Anonymization

The exposure of information about patients and their health may lead to privacy issues. Information privacy is defined as the desire of individuals to control or have some influence over data about themselves (Bélanger & Crossler, 2011). It is, in other words, the right of individuals to determine how and to what extent information is communicated to others. Protecting data privacy can be done either by restricting access to the data by using control methods or by anonymizing the data. The main idea is to publish sensitive data for gaining valuable insights without questioning an individual’s privacy. This approach is called privacy-preserving data publishing (PPDP) (Victor & Lopez, 2016).

Usually, a data publishing scenario consists of three main stakeholders namely, the owner of the data who has collected/created it, the holder of the data, and finally the recipient of the published data who will use it. In the most basic form of PPDP, the data publisher has a table of the form that contains an explicit identifier, quasi identifier, sensitive attributes, and non-sensitive attribute. An explicit identifier refers to a set of attributes that uniquely identify the record owner such as name, address, and national identity number. A quasi identifier refers to a set of attributes that could potentially identify the record owner such as age, sex, and zip code. Sensitive attributes consist of sensitive, specific information such as disease, salary, and disability while non-sensitive attributes are the non-sensitive information that does not fit into any of the three previously mentioned categories (Victor & Lopez, 2016).

During the data publishing phase, explicit identifiers get removed from the published dataset, only the quasi-identifiers, sensitive attributes, and non-sensitive attributes are published. However, the published dataset undergoes modification processes to make it anonymized before it is published to the recipient. The modification processes are accomplished by performing a variety of anonymization operations on the dataset (Xu, Ma, Tang, & Tian, 2014). Anonymization is defined as a technique that uses data distortion to preserve the privacy of public data to be published (Sharma, Jayashankar, Banu, & Tripathy, 2016).

It refers to the PPDP approach that aims at hiding the identity and/or the sensitive data of record owners to prevent from linkage attack assuming that sensitive data must be retained for data analysis (Fung, Wang, Chen, & Yu, 2010). In (Fung, Wang, Chen, & Yu, 2010), the authors have identified and summarized anonymization approaches into three main operations which are presented in figure 2.1.

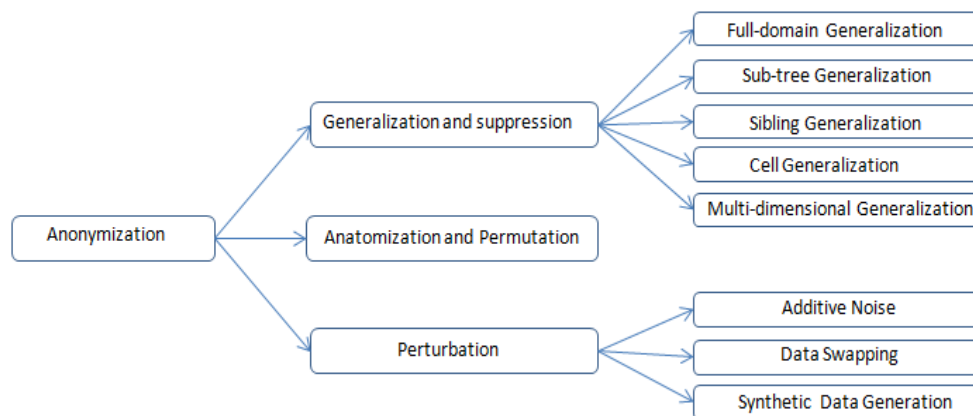


Figure (2.1) Anonymization Approaches (Fung, Wang, Chen, & Yu, 2010)

The main objective of the generalization approach is the replacement of specific values with more general ones, and as a result, many tuples of the data will be having the same set of values for quasi-identifiers. In the Anatomization and permutation approach, the main goal is to de-link the relation between quasi-identifiers and the sensitive attributes of the record owner, while the perturbation approach aims at adding noise to the original dataset before it is received by the user/recipient.

The privacy of the individual's published data can be breached by two common attacks namely; Probabilistic attack and Linkage attack (Rashid & Yasin, 2015). The probabilistic attack happens when the attacker makes successful speculation by inferring the potential fit between the individuals and randomized records. A Linkage attack happens when an attacker becomes

able to identify an individual from the published data. Linkage attack, as the name suggests, tends to link an individual to a record or value in a given table. There are three types of linkage attacks described in (Manta, 2013) namely; record linkage, attribute linkage, and table linkage. The privacy breaches on published data can be categorized into three main types which are identity disclosure, sensitive link disclosure, and sensitive attribute disclosure (Liu, Das, Grandison, & Kargupta, 2008)(Zheleva & Getoor, 2007). The identity disclosure attack happens by exposing the record owner (individual) leading to the revelation of information of the user and relationship he/she shares with other individuals. Sensitive link disclosure happens when the associations between two individuals are revealed, while the sensitive attribute disclosure attack happens when an attacker obtains the information of a sensitive and confidential user attribute to link it with an entity. Such attacks create the challenge of maintaining the privacy of individuals while making their data accessible to their full potential.

Various privacy models aim at preventing linkage attacks on published datasets. These privacy models ensure privacy either at the record level, attribute level, table level, or at all levels of the data published. Each privacy model employs one or more of the anonymization operations for giving better results. This research will focus on studying the privacy models identified in (Victor & Lopez, 2016) that can be extended to the big data domain namely; k-anonymity, l-diversity, and t-closeness privacy models.

2.2.1 K-anonymity

The K-anonymity model is defined as a property possessed by certain anonymized data. It is a model proposed in (Sweeney, 2002) as an attempt to address the problem “how can data holder release a version of its private data with scientific guarantees that the individual cannot be re-identified while the data remain practically useful?” For example, a data holder such as a medical institution may want to release a table of medical records. Even though the names of the individuals can be replaced with dummy identifiers such as number or code, some set of attributes (quasi-identifier) can cause leakage to confidential information such as the date of birth, zip code, and the gender in the disclosed table which can uniquely determine an individual. K-anonymity employs both generalization and suppression techniques. Attributes in k-anonymity are suppressed or generalized until each row is identical with at least k-1 other rows. At this point, the database is said to be k-anonymous which is not prone to definite database linkages. At worst, the data released as a response to an individual’s entry can be narrowed down to a group of k individuals with guaranteed accuracy. For example, it is not

possible to identify a man in a released table if the information available is only the gender and the date of birth. There are k men who have the same date of birth and gender.

However, k -anonymity has certain drawbacks that make it less efficient in several cases. There are two major attacks known as Homogeneity and Background Knowledge attacks (Maheshwarkar, Pathak, & Choudhari, 2012).

Homogeneity Attack happens due to the lack of diversity in the sensitive attributes (Hussien, Hamza, & Hefny, 2013). Suppose A is intending to infer B's medical status in a particular table. A knows B's ZIP Code 12345 and his age is 35. Using this knowledge, A can know that B's records fall in a certain range 9, 10, 11, 12 who suffer from cancer concluding that B has cancer.

Background Knowledge attack happens when an attacker knows background knowledge and uses it to eliminate possible values for the sensitive attributes of the victim. For example, the attacker knows that Alice is 35 years old, female, writer, and has been to the hospital which published the table. The attacker can see that all the female writers of age 35 suffer from a common disease which is HIV. The attacker can then conclude that Alice suffers from HIV disease. This attack is known as a positive disclosure attack (Manta, 2013).

2.2.2 *l*-diversity

l-diversity is a form of group-based anonymization that is used to preserve privacy in data sets by reducing the granularity of data representation (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2006). It is the model used in the proposed system design in this thesis (Chapter 5). *l*-diversity is an extension of the k -anonymity model that aims at handling some of the weaknesses in the k -anonymity model. A Quasi-identifier block Q is *l*-diverse if it contains at least l well-represented values for each sensitive attribute S . The table T is *l* diverse if every Q block (Equivalence class) is *l*-diverse (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2007). If there are at least l well-represented values for sensitive attributes, the adversary needs to eliminate $l-1$ possibilities of sensitive attributes to gain a positive disclosure about the information of the individual.

The main principle of the *l*-diversity model is to have a diversity of the sensitive attributes within each quasi-identifier equivalence class. Each equivalence class has at least l well-represented sensitive values. This overcomes the drawback of the k -anonymity model if a particular equivalence class lacks diversity which enables attackers to perform linkage attacks.

However, the *I*-diversity model also suffers from major drawbacks, it is difficult and often unnecessary to achieve. If the sensitive attribute is just taking one of two values ‘affected’ or ‘not affected’ and if 90% of the people are in the category ‘not affected’ then it may be acceptable for the individuals in that category to reveal their status. But this will not be the case for individuals who were tested as positive. They always want to keep their information private. This challenge is overcome by advancement on the model which is further explained in Chapter 5.

I-diversity model is prone to two main types of attacks; the first attack is a Skewness attack which can take place if each block of quasi-identifiers (equivalence class) has an equal probability for positive and negative values of sensitive attributes. The second attack is the Similarity attack which happens when the values of sensitive attributes look different but have the same or common meaning. This attack happens due to the principle of the model which considers the diversity of sensitive attributes but does not consider the closeness of various values in the sensitive attributes meaning wise.

2.2.3 *t*-closeness

The *t*-closeness model proposed in (Li, Li, & Venkatasubramanian, 2007) as a further refinement of the *I*-diversity anonymization that is used to preserve privacy in data sets by reducing the granularity of data representation. The *t*-closeness model extends the *I*-diversity model by treating the values of an attribute distinctly by considering the distribution of data values for that attribute. *t*-closeness the model seeks to limit the amount of information that an adversary can obtain about the confidential attribute of any specific subject/individual. To this end, *t*-closeness requires the distribution of the confidential attributes within each of the equivalence classes to be similar to their distribution in the entire data set. The main principle of the *t*-closeness model is that the distribution of sensitive attributes within each quasi-identifier group should be “close” to their distribution in the entire original database. An equivalence class is said to have *t*-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold *t*. A table is said to have *t*-closeness if all equivalence classes have *t*-closeness.

However, in (Victor & Lopez, 2016), several major issues with *t*-closeness are outlined. One of the main issues with *t*-closeness is that different levels of sensitivity should be specified for different sensitive attributes; *t*-closeness prevents attribute disclosure, but it does not prevent identity disclosure. Another issue with *t*-closeness is that the more sensitive attributes published

in a table the more privacy is being questioned. Moreover, the quality of data after performing the anonymization processes such as generalization and suppression is affected by the t -closeness approach.

2.2.4 Information privacy threats

The main threat to the privacy of patients' information when it is available for research is the re-identification of individual patients. Privacy threats relate to three types of attributes in datasets which are explicit identifiers, quasi-identifiers, and sensitive attributes. Explicit identifiers are the attributes that can be used to directly identify a patient, such as a name, email address, phone number, physical address ... etc. Quasi-identifiers are attributes that -when combined- can lead to identity disclosure, such as patients' demographical information which includes patient's date of birth, gender, zip code ... etc (Xiao & Tao, 2006). Sensitive attributes are information that patients do not want to disclose or be associated with such as medical conditions (e.g. cancer, HIV, or psychiatric conditions). The privacy of the individual's published data can be breached by a Probabilistic attack and Linkage attack (Rashid & Yasin, 2015).

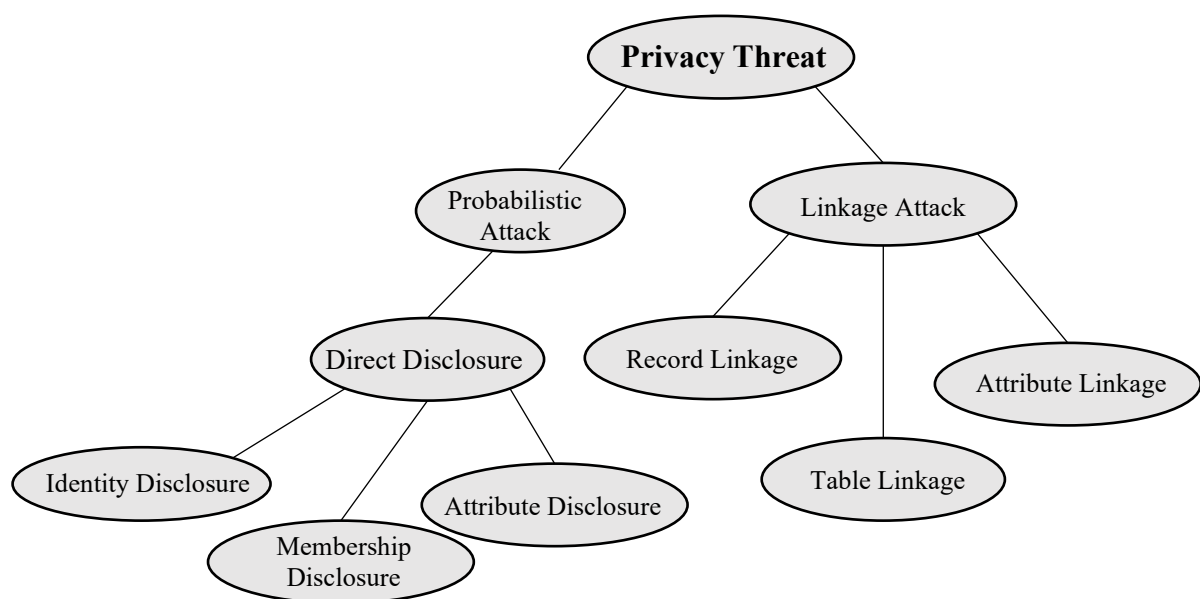


Figure (2.2) Privacy threats tree

The probabilistic attack happens when the attacker makes successful speculation by inferring the potential fit between the individuals and randomized records. Linkage attack, as the name suggests, is that the adversary reveals sensitive information by the means of linking. There are three types of linkage attacks described in (Manta, 2013): record linkage, attribute linkage, and

table linkage. Adversaries in linkage attacks link anonymized datasets with other datasets obtained from different sources such as the government voting dataset (Rashid & Yasin, 2015).

The authors in (Gkoulalas-Divanis, Loukides, & Sun, 2014) identified three main privacy threats that need to be mitigated to assure the privacy of patients' information when aggregated for research purposes namely identity disclosure, membership disclosure, and attribute disclosure. Identity disclosure is also called re-identification which occurs when an attacker becomes able to associate a patient with their information in a published dataset (Sweeney, 2002). For example, an attacker can identify Ray Gather in Table (5.5) even after removing his explicit identifiers (name and mobile number), because he is the only one in the table who was born on 22-01-1981 and lives in zip code 49511.

A membership disclosure attack happens when an attacker can conclude with high confidence that an individual's information is contained in the published dataset. For example, if a dataset contains information only about positive HIV is published, the existence of patient records in the dataset reveals that the patient was diagnosed with positive HIV (Nergiz, Atzori, & Clifton, 2007).

Explicit identifiers		Quasi identifiers			Sensitive attribute
Name	Mobile Number	Date of Birth	Zip code	Gender	DNA
Steve Jordan	96552530	15-05-1979	44985	Male	TG ... T
Ray Gather	95978813	22-1-1981	49511	Male	TT ... A
Hannah Joe	95225884	30-09-1983	44985	Female	CG ... G
Andrew Keene	95225884	27-07-1990	49781	Male	AT ... G
Julia Robin	95225884	16-06-1988	49771	Female	TA ... G

Table (2.1) Example of types of attributes in a relational table

Attribute disclosure is an attack in which an individual patient is associated with information related to their sensitive attributes (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2006). It aims at identifying the individual by combining released records with background knowledge (Abid, Malik, Usman, Hasan, & Khalid, 2018). For example, if the sensitive attribute is the cost of hospitalization, this may indicate the nature of the treatment required. If it is high, it may reveal that the patient required hospitalization for a rare disease or relatively costly to treat conditions.

2.3 Cloud Computing Solution

The technology of cloud computing represents a different method for remotely managing and architecting computer resources (Winans & Brown, 2009). Cloud computing services are delivered through a network which is usually the internet. It is a computing model that enables ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released without a need for great management efforts or service provider interaction (Mell & Timothy, 2011). Cloud computing facilitates a computing-as-a-service model where computing resources are made available as a utility service. It allows the convenience of using as many resources as demanded by the user in a pay-as-you-go basis which makes it different from the earlier computing models in which enterprises have to invest enormous funds to implement and build their own IT infrastructures (Mishra, Das, Kulkarni, & Sahoo, 2012). It facilitates the possibility for users to rent only at the time of need the desired amount of computing resources out of a huge mass of distributed computing resources without worrying about the locations or internal structures of these resources (Kuribayashi, 2012). Cloud computing leverages the virtualization of computing resources aiming at allowing customers to provision resources on-demand (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2008). Today, cloud computing technology is regarded as an important trend towards future's distributed and ubiquitous computing services offered over the global internet (Pedersen, et al., 2011), it is also gaining a great deal of attention and popularity in our current society due to the benefits that it can flexibly offer to its users with various applications for various purposes within the context of a pay-as-you-go model.

The technology of cloud computing offers services in three primary models described in (Dialogic, 2010) namely: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). IaaS is a cloud-based service model that enables users to control and manage applications, storage, network connectivity, and operating systems without having control over the cloud architecture (Dialogic, 2010). It is a model that aims to provide computing infrastructures such as servers, storage, network, and operating systems that are distributed as a measured, scalable service (Kepes, 2011). Such a model benefits organizations/users by avoiding the expense associated with the ownership and management/maintenance of such computing systems. It generally includes numerous users sharing the capabilities of a single piece of computing hardware. PaaS is a cloud-based model that enables users to access platforms and deploy their own applications on these platforms.

PaaS allows the creation of web applications easily without the complexity of purchasing and maintaining the infrastructure underneath them (Kepes, 2011). In PaaS, users create their own software using tools and libraries provided by cloud providers, but they are not permitted to manage or control the underlying cloud infrastructure such as network, servers, operating systems...etc., however, they are able to control their deployed applications and have access to the configuration settings for the application-hosting environment. SaaS is an application delivery model that allows users to utilize a software solution over the internet (Mell & Timothy, 2011). It is a cloud-based model that allows accessing an application that is hosted in a remote datacentre via an internet connection. In SaaS, users can purchase the accessibility and usability of an application or service that is hosted in the cloud. In the SaaS service model, applications are accessible from various devices through user interfaces. Users are not able to manage or control the underlying cloud infrastructure such as networks, servers, and operating systems, however; there is a possible exception of limited user-specific application configuration settings (Mell & Timothy, 2011). Facebook is an example of a SaaS service; Facebook users can create and access their social/commercial accounts on the Facebook site through any internet-enabled devices while the service is hosted in a remote datacentre.

In terms of business, cloud computing technology is seen more as a new business model rather than a new technology. The technology grants opportunity for acquisition and management of computing assets and software platforms and capabilities for the prompt addition of new features concerning business changing needs. Cloud computing technology enables companies to perform their main functions in an environment that offers a good basis for starting or expanding a business without big investments (Kiryakova, Angelova, & Yordanova, 2015). The National Institute of Standards and Technology (NIST) points out the 5 main characteristics of cloud computing technology that distinguish it from other technologies which are On-demand self-service, Broad network access, Resource pooling and sharing, Rapid elasticity, and Measured service (Mell & Timothy, 2011) (Spinola, 2009).

The on-demand self-service characteristic enables users to unilaterally declare and obtain computing resources. It allows cloud users to individually provision computing capabilities as needed without a need for human interaction with each service provider. Cloud computing creates the illusion of infinite computing resources available on-demand and eliminates the need to make preliminary plans for users' long-term supply. Permanent and broad network access characteristic allows the availability of capabilities over the network which can be accessed through standard mechanisms that promote use by heterogeneous thin or thick client

platforms such as mobile devices, laptops, and workstations. Computing resources are available anytime and anywhere over the network via standard mechanisms which are the Web protocols. Resource pooling and sharing characteristics allow multiple users to be served using a multi-tenant model. In resource pooling, users are serviced with different physical and virtual resources which are dynamically assigned and reassigned according to users' requirements. Those resources may include memory, storage, and bandwidth. Rapid elasticity characteristic refers to the capability of delivering services at any time and quantity according to users' requirements. Depending on the current needs, users can dynamically increase or decrease the rented computing resources according to their needs. Measured service characteristic allows the feasibility of measuring and controlling the computing resources usage by leveraging a metering capability at some level of abstraction appropriate to the type of services such as storage and processing. Payment for cloud services depends on consumption. Cloud services are offered to users in a pay-per-use manner. Enterprises using cloud computing technology for their businesses report up to 30% economic saving along with other related benefits such as more effective mobile working, higher productivity, or the standardization processes (Bradshaw, Folco, Cattaneo, & Kolding, 2012).

The concept of cloud computing is divided into three forms of cloud deployments namely: private cloud, public cloud, and hybrid cloud (Aggarwal, 2018). All deployments have significant characteristics; however, to obtain the benefits of cloud computing technology, identifying the requirements that need to be addressed is essential.

A private cloud is a deployment model of cloud computing that is provisioned solely for a single client which is usually an organization. A private cloud may be owned, operated, and managed by its client or a third party. Private clouds can be hosted internally or externally. The main objective of implementing a private cloud is avoiding security issues as it is implemented safely with private firewalls and other means of security, as well as promoting better efficiency in determining workload and usage priorities (Owopetu, 2013).

Public cloud as the name suggests is a cloud model that is provisioned for public use. It is a computing infrastructure that is hosted at the vendor's premises. The computing infrastructure is shared between organizations. The reason for its name "public" is because it is meant to be accessed by various users from the public. Public clouds are owned by various types of organizations, such as business organizations, academic organizations, government organizations, or combinations of them (Vikas, Gurudatt, Vishnu, & Prashant, 2013).

Hybrid cloud is a combination of both private and public implementations that gives business entities the advantage of both cloud environments (Vikas, Gurudatt, Vishnu, & Prashant, 2013). Hybrid clouds are unique entities, bound together by standardized technology that allows for data and application portability and interoperability. For example, an organization wishes to share its services and products with its clients across internationally, but at the same time wants to hide the confidential information from them, Hybrid cloud architecture would be a solution for such organizations (Aggarwal, 2018). Hybrid clouds can be hosted internally and externally.

2.3.1 Barriers to adopting cloud computing

Since the cloud computing phenomenon was introduced, there has been an unceasing interest in research across the globe. Cloud computing opens doors to multiple, unlimited venues from elastic computing to on-demand provisioning to dynamic storage and computing requirements fulfillment. However, despite the potential gains achieved from technology, there are still several challenges that hinder its adoption. One of the most significant challenges in the adoption of cloud technology is security, followed by issues related to compliance, privacy, and legal matters (Hashizume, Rosado, Fernández-Medina, & Fernandez, 2013). Because cloud computing is a relatively new computing model, there is a significant deal of uncertainty related to how can security be achieved at all levels such as network, host, application, and data levels. Organizations and individuals are often concerned about how security and compliance integrity can be maintained in this new computing model. Such concern has led the organization to hesitate to move critical resources to the cloud (Rosado, Gómez, Mellado, & Fernández-Medina, 2012).

In a cloud environment, users outsourcing their data and applications can only rely on the cloud service provider (CSP) to protect the security of their data and applications. The security concerns here rise due to the fear of the unknown. Sharing computing resources among multiple users generate a risk of data misuse. The authors in (Rao & Selvamani, 2015) highlighted data-related security challenges in a cloud-based environment. The authors outline three main areas of data security namely: confidentiality, integrity, and availability. Enhancing the security of data stored on the cloud requires sufficient authentication and authorization mechanisms to assure sufficient access control to it, as well as guaranteeing the availability of the data whenever it is needed by its owner. The lack of users' control over their data on the cloud raises extensive privacy concerns since the sensitive information of cloud users may be accessed and

analyzed by unauthorized parties (Liu, Sun, Ryoo, Rizvi, & Vasilakos, 2015). The protection of user privacy and multimedia data/application secrecy from an adversary is key to establish and maintain consumer's trust in a cloud platform (Al-Qurishi, et al., 2018). Regulation compliance is another challenge that is hindering the wide adoption of cloud computing (Yimam & Fernandez, 2016). It implies enforcing the rules that implement the policies defined in the regulations. Regulations are sets of policies that govern the use of sensitive business data. According to the National Institute of Standards and Technology (NIST), organizations are responsible for compliance-related issues, and not being compliant to regulations may result in penalty fees, lawsuits, and bad business reputation (Yimam & Fernandez, 2016). The main goals of these regulations are to protect the security and privacy of consumers' information by enforcing attributes such as confidentiality, integrity, availability, and accountability.

Various regulatory bodies defined rules and regulations to ensure the security of data and permit disclosure under acceptable circumstances. Such regulations involve a wide range of applications and practices listed and explained in (Khan, 2016) which include common criteria, trusted computing, and privacy acts. When organizations have their data and workload processed in-house, they usually have ultimate control over their sensitive data, but when such responsibilities are outsourced to the cloud, organizations require verifying that the cloud service providers respect the regulatory and compliance requirements, especially when these organizations belong to critical domains such as government, finance, or healthcare. Organizations that belong to critical domains are required to adhere to specific regulations such as Health Insurance Portability and Accountability (HIPAA), the Payment Card Industry Data Security Standards (PCI DSS), and the Federal Information Security Management ACT (FISMA). General concerns in this context include the need for consent from users when dealing with personal data, the need for strong access control mechanisms, compliance to data jurisdictions, and compliance to data confidentiality regulations (Cloud Security Alliance, 2011). Therefore, the possibility of a lack of enforcement of security regulations is considered an obstacle on the way to adopting cloud services (Phaphoom, Wang, Samuel, Helmer, & Abrahamsson, 2015).

2.3.2 Cloud computing for healthcare

Healthcare is an important pillar of society, critical for effectively responding to public health emergencies, and addressing disease, ill health, and poverty brought on by communicable disease and non-communicable disease and cancer (Atun, 2012). The unceasing demand for

cost-effective, time-effective, and preventive healthcare is forcing radical changes in current healthcare systems, requiring them to take full advantage of the capabilities of modern technology including information technology (Christodoulakis, Asgarian, & Easterbrook, 2017). Medicine is an increasingly data-intensive and collaborative endeavor. In the past century, technology has played a critical role in defining, driving, and reinventing procedures, devices, and pharmaceuticals in the healthcare sector. The need for adequate resources to process, store, exchange, and use large quantities of medical data has brought the attention of researchers to cloud computing. Cloud computing technology has been introduced only recently but is already one of the major topics of discussions in research and clinical settings (Kagadis, et al., 2013).

Kuo (2011) recognized the technology of cloud computing as a potential solution for enabling healthcare organizations to focus their efforts on clinically relevant services and improved patient outcomes. The term “Cloud Computing” is a new name for an old concept; the delivery of computing services from a remote location, analogous to the way electricity, water, and other utilities are provided to most customers (Fischer & Figliola, 2013) (Tebaa, Hajji, & Ghazi, 2012). Cloud computing appears to be the dreamed vision of the healthcare industry, it matches the need for healthcare information sharing directly to various healthcare-related parties over the internet, regardless of their location and the amount of data being shared (Guo, Kuo, & Sahama, 2012).

Today, cloud computing is making its way in many fields in the healthcare domain due to the benefits that technology grants such as minimum cost, effective use of resources and maximized availability of services. The accelerating adoption of cloud computing in the healthcare domain represents a change in the way information technology is sourced (Dubey & Vishwakarma, 2016). Cloud technology is used to create connecting networks between healthcare institutions, healthcare practitioners, and patients by providing applications, services, and storage of data in the cloud. Among the reasons for the interest of the healthcare domain in cloud computing is the need for collaboration among the increasing number of remote and mobile workers, several office locations, a desire to improve patients quality of service and even present goals of improving operational excellence with lower cost (Lester, Boateng, Studeny, & Coustasse, 2016).

However, despite the attraction towards adopting cloud computing in the healthcare domain, like all other fields, the healthcare industry is still hesitant to embrace the technology due to

concerns related to data security such as privacy, availability, and integrity (Dubey & Vishwakarma, 2016). Organizations in the healthcare domain hold sensitive data that should never be disclosed to unauthorized users as protection to patients' privacy. In healthcare, because of the probable disclosure of medical records stored and exchanged on the cloud, the patients' privacy becomes vulnerable (Zhang, et al., 2017) (Jabbar & Najim, 2016) (Chauhan, Sanger, & Verma, 2015).

In (Hu & Bai, 2014), the authors have conducted a systematic review of computing in eHealth. The goal of the review was to identify the state of the art regarding the adoption of cloud computing in the healthcare domain, the intention was to pinpoint challenges and possible directions for researchers and application developers based on the current literature. It was found in the study that the application of cloud computing in the healthcare domain was still immature. Adhering to legal frameworks with regards to storing and sharing healthcare information in privacy-preserving manners was found one of the issues hindering the adoption of cloud computing technology for the healthcare industry. As a result of the review, the authors found that a hybrid cloud model that contains access controls and security protection techniques would be a reliable solution. The authors proposed that hospitals and healthcare centers keep their data in private clouds, and patients' daily self-management data could be published in a confident public cloud, and patients should decide who can access their data and conditions for sharing it.

Another systematic review was conducted in (Mehraeen, Ghazisaeedi, Farzi, & Mirshekari, 2017). The authors reviewed and covered articles published between the year 2000 to the year 2015. The main focus of the review was on the security issues of cloud computing in the healthcare domain. The authors found that security and privacy issues have played the most important role in hindering the acceptance of cloud computing technology for healthcare. Issues such as identity management and access control, authentication and authorization, and cybercriminals were identified in the survey as the major security issues identified the healthcare cloud computing.

Similarly, the authors in (Jain & Singh, 2017) have conducted another systematic review which included 51 published articles from the year 2014 to 2017. The survey was based on security challenges in healthcare analysis over the cloud. The goal of conducting the survey was to investigate the challenges in cloud computing related to healthcare, it included a detailed review of the healthcare cloud computing security and privacy issues and explored the main

challenges with a focus on the compliance concerns and ensuring trust data security. Various approaches to preserve the privacy of health information in the cloud environment were outlined in the survey. However, the authors concluded with a list of security and privacy-related challenges that need to be addressed before obtaining the best of what cloud computing can offer to the healthcare industry. These challenges were: Data Security, Access Control, and Protection from malicious Code. Data security is a major issue in cloud computing. It involves a number of aspects associated with it such as privacy, confidentiality, integrity, reliability, availability, backup, and recovery. Overcoming such a challenge is vitally important before cloud computing offers its benefits to the healthcare domain. The security of information refers to preserving information and information systems from unauthorized access, use, disclosure, modification, destruction, or interference. The general requirements of cloud security depend on many issues which include privacy, trust, integrity, availability, and confidentiality. Access control to the data stored on the cloud is another challenge that needs to be addressed. Digital identity management is crucial in cloud computing architectures to authenticate users and support flexible access control (Elisa Bertino, 2009). Normally, the owner of the data creates a set of access control rules on their data and send the data along with the access control policy. Users can view or use the data only if the access control policy set by the data owner allows. However, a member of the owner's panel would still be allowed to access the data. Access control policy should lock without the permission of the owner. This is considered a challenge in cloud computing data security. Malicious code is a code created by hackers to alter information. It is an application security threat that cannot be efficiently controlled by conventional antivirus software alone. In healthcare cloud environments, individuals may monitor the sequence of events to obtain unauthorized access to sensitive information. Therefore, developing mechanisms to deploy efficient auditing and accountability that anonymously monitor the utilization of health records is considered highly important towards better adoption of the technology in the healthcare domain.

Security and privacy issues create a barrier that hinders the adoption of cloud computing technology for healthcare information systems. Without achieving sufficient mechanisms to assure the security and the privacy of patients' sensitive records, the adoption of cloud computing in the healthcare domain will remain limited. The security of information refers to preserving information and information systems from unauthorized access, use, disclosure, modification, destruction, or interference. The general requirements of cloud security depend on many issues which include privacy, trust, integrity, availability, and confidentiality. In this

regard, the authors in (Sangeetha & Kavitha, 2016) outline the security-related requirements that need to be met in cloud computing for storing and sharing healthcare information:

- Data owners should be able to assign other cloud users with different access privileges to their data.
- The cloud needs to be able to support dynamic requests so that data owners can add or revoke access privileges to other users allowing them to create or delete their data.
- Users' privacy must be protected against the cloud so that they can conceal their private information while accessing the cloud.

Concurrently, the authors in (Raval & Jangale, 2016) proposed a cloud-based system diagram that can satisfy the need of sharing healthcare information concerning the data-related security measures such as privacy, confidentiality, access control, and prevention from malicious attacks. The proposed system provided an environment in which patients' records are stored and referenced by medical practitioners for healthcare services purposes. The following includes the main characteristics of the proposed system:

- Medical practitioners can access patients' records dataset using a unique registration number associated to them or to their affiliated institutions who are granted access permission. Any user that accesses the database must be registered with a license number and allowed to access the data.
- Patients' records are stored in the database with an identification number for each individual patient which is generated when the patient is first registered in the system.
- Patients are granted Read-Only privilege on their data that is stored on the system.
- Hospitals or medical practitioners can update any patients' records using patients' identification number and their license numbers.

However, to date, achieving the above implementation of the cloud system for the healthcare domain remains elusive. The complexity of healthcare data and the variety of healthcare practitioners in terms of their roles in the sector make it difficult to maintain sufficient access control which may lead to questioning the security of patients' records stored on the cloud.

2.3.3 Privacy-preservation approaches

Due to the great interest in the adoption of cloud computing in the healthcare domain, there have been tremendous efforts found in the literature that aimed to address the challenges related to information privacy. Organizations in the healthcare domain hold sensitive data that requires

a high level of privacy protection (Zhang, et al., 2017). There are two types of privacy-preserving mechanisms identified in the literature for preserving the privacy of healthcare information on the cloud namely: cryptographic mechanisms (Claret, 2011) and non-cryptographic mechanisms.

In cryptographic mechanisms, encryption techniques are employed such as symmetric key encryption (Yassein, Aljawarneh, Qawasmeh, Mardini, & Khamayseh, 2017), public-key encryption (Abdalla, Benhamouda, & Pointcheval, 2016), and alternative cryptographic primitives which are explained further in this section, while the non-cryptographic mechanisms include access control mechanisms such as Role-Based Access Control (RBAC) (Bertino, Bonatti, & Ferrari, 2001), Attribute-Based Access Control (ABAC) (Hu, Kuhn, & Ferraiolo, 2015), Mandatory Access Control (MAN) and Identity Based Access Control (IBAC) (Osborn, Sandhu, & Munawer, 2000)...etc. A comprehensive systematic review of cloud information security and privacy-preserving approaches that are in the literature was conducted in (Chenthara, Ahmed, Wang, & Whittaker, 2019). The review aimed to investigate the security and privacy requirements of smart health data in the cloud arena, summarize a brief architecture of e-Health clouds using taxonomy over privacy-preserving approaches, and finally discuss the merits and drawbacks of the furnished mechanisms to indicate future research directions.

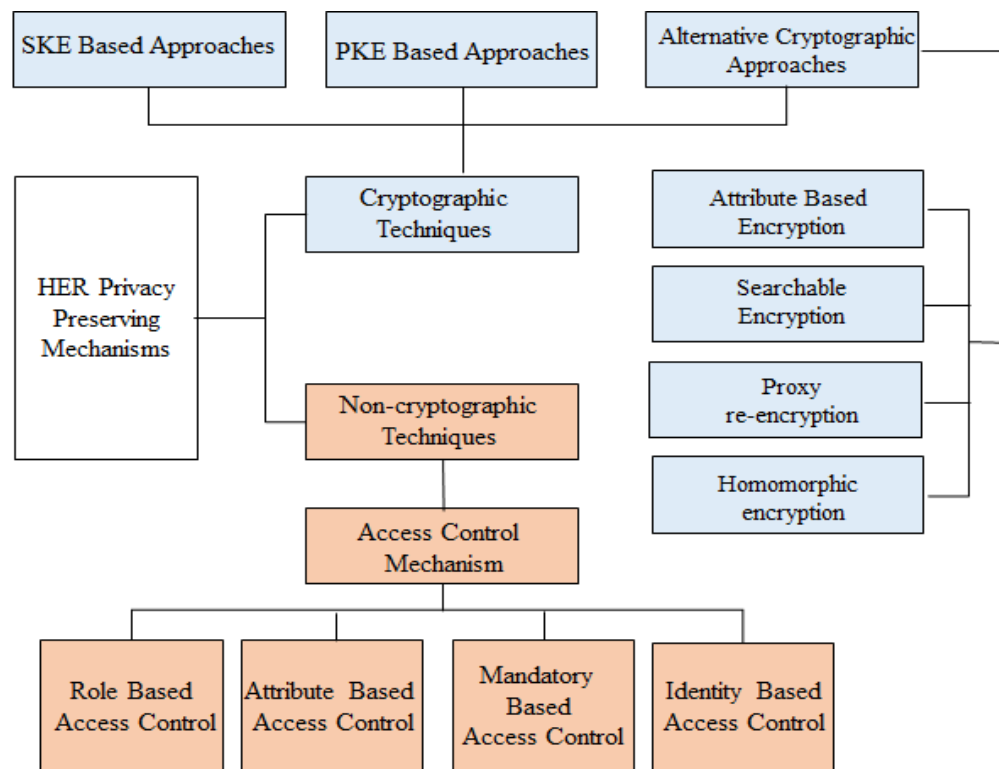


Figure (2.3) Classification of Privacy Preserving mechanisms in electric health records (Chenthara, Ahmed, Wang, & Whittaker, 2019)

As seen in Figure (2.3), the privacy-preserving mechanisms are classified: cryptographic techniques and non-cryptographic techniques. Cryptography refers to the science of using mathematics to encrypt and decrypt data. It enables to store sensitive information or transmit it across insecure networks so that it cannot be read by anyone except the intended recipient (Barakat, Eder, & Hanke, 2018). The main objectives of cryptography are to achieve confidentiality, integrity, non-repudiation of information, and authentication of the sender and receiver in contexts of sending and receiving messages (Rouse, 2019).

Cryptography approaches include symmetric key encryption (SKE) and asymmetric key encryption which is also known as Public Key Encryption (PKE). In SKE, a single shared key is used for both encryption and decryption processes, and it is considered highly effective when used for electronic health record systems. SKE has the unlinkability characteristic which is highly important to protect the privacy of patients' information. The numbers of patients' electronic medical records, generated using the SID in the health data card, random value, and treatment serial number, are all different even for the same patient, therefore, the unlinkability characteristic of SKE makes effective in terms of security and privacy of information. However, the use of SKE introduces additional complexity that cannot be avoided in the system due to the need for additional access control mechanisms to effectively share information in the healthcare (Li, Chang, Huang, & Lai, 2011). There are a number of commonly used SKE-based algorithms found in the literature such as Advanced Encryption Standard (AES) (Abdullah, 2017), Blow Fish (Gowda, 2016), and Data Encryption Standard (DES) (Schneier, 2015).

The PKE approach requires two different keys; public key and a private one. In PKE, the senders' private key or the receiver's key, or both can be used by the sender according to the cryptographic function (Mohamed & Harb, 2015). The authors in (Stallings, 2005) categorize public key cryptography into three different categories namely encryption/decryption, digital signature, and key exchange. In encryption/decryption, the message is encrypted with the recipient's public key and decrypted with the recipient's private key. In a digital signature, the sender uses its private key to sign a message, while in the key exchange, a session key can be exchanged between two sides of communication with some type of cooperation. The downfall of PKE schemes is that they are computationally inefficient when they are used on their own due to the large size of keys. PKE schemes perform more efficiently when they are in combination with SKE schemes (Chenthara, Ahmed, Wang, & Whittaker, 2019).

Cryptographic primitives are another approach for protecting the privacy of healthcare data in cloud environments. Cryptographic primitives (Lazar, Chen, Wang, & Zeldovich, 2014) are well-established, low-level cryptographic algorithms that are frequently used when building cryptographic algorithms for computer security systems. Cryptographic algorithms provide means of security such as confidentiality, integrity, and authentication based on a solid mathematical foundation that prevents against powerful adversaries like the NSA (Schneier, 2013) (Snowden, 2013). Attribute-Based Encryption (ABE), Proxy Re-Encryption (PRE), Homomorphic Encryption, and Searchable Encryption (SE) are all examples of cryptographic primitive approaches.

Attribute-Based Encryption (ABE) is an approach that was introduced in 2005 by Amit Sahai and Brent Waters (Sahai & Waters, 2005). The approach is based on public-key encryption to protect cloud data where the encryption and decryption are on the bases of user attributes. The encryption in the ABE approach is based on the access-structure policy in which the only way of decrypting a ciphertext is to have the attributes of the user match with the attributes of the ciphertext. There are two main types of ABE which are Cipher Text Policy Attribute-Based Encryption (CP-ABE) described in (Li, Yu, Zheng, & Ren, 2013) and Key Policy Attribute-Based Encryption (KP-ABE) described in (Bethencourt, Sahai, & Waters, 2007). In KP-ABE, the policy to access the text is enciphered in the user's secret key, and the decryption of the ciphertext only happens when the user attribute matches with the access policy. In CP-ABE, the private key of each user is associated with a set of attributes, and the ciphertext is associated with a universal set of attributes. The ciphertext gets encrypted only when the user attributes match the access policy.

Searchable Encryption (SE) is a cryptographic tool that enables to search through a set of data or a string or a file that contains a specific keyword. It is an approach that enables to search through data while it is encrypted. Searchable encryption schemes enable the receiver to search an email containing a particular keyword among a set of emails in his/her account (Pramanick & Ali, 2017). There are two types of searchable encryption namely symmetric searchable encryption (Song, Wagner, & Perrig, 2000) and asymmetric searchable encryption (Kamara, 2010). There are many searchable encryption schemes found in literature such as Public Key Encryption with Keyword Search (PEKS) which was proposed in (Boneh, Crescenzo, Ostrovsky, & Persiano, 2004), Extension of PEKS scheme proposed in (Abdalla, et al., 2005), Deterministic and Efficiently Searchable Encryption (DESE) proposed in (Bellare, Boldyreva, & O'Neill, 2007), Symmetric Searchable Encryption (SSE) (Curtmola, Garay, Kamara, &

Ostrovsky, 2006), Multi-Keyword Fuzzy Search (Wang, Yu, Lou, & Hou, 2014), and Scoring and Ranking which was proposed in (Orencik, Selcuk, Savas, & Kantarcioglu, 2016). The authors in (Pramanick & Ali, 2017) have conducted a comparative survey on various techniques of searchable encryption schemes. The conclusion reached in the survey was that searchable encryption schemes still have some limitations related to complex Boolean queries and their implementation models.

Proxy Re-encryption is another cryptographic approach that permits the semi-trusted proxy server to re-encrypt the ciphertext, which is encrypted by one user's public key, into another ciphertext (Blaze, Bleumer, & Strauss, 1998). For example, Alice wants to send a message to Bob, she sends the message (M) to Bob through a semi-trusted proxy server without sharing Alice's private key to either Bob or the proxy server, and without disclosing the secret message to the proxy. The proxy re-encryption scheme in this example converts a ciphertext for Alice into a ciphertext for Bob without reading the secret text that is encrypted, the proxy only requires a re-encryption key from Alice to achieve that.

Homomorphic encryption is defined as a form of encryption method that allows specific types of computation to be carried out on ciphertexts and generates encrypted results which, when decrypted, matches the result of operations performed on the original texts (Yi, Paulet, & Bertino, 2014). Homomorphic encryption refers to a class of encryption methods envisioned by Rivest, Adleman, and Dertouzos in 1978 (Rivest, Shamir, & Adleman, 1978), and was first constructed by Craig Gentry in 2009 (Gentry, Sahai, & Waters, 2013). In the context of conventional symmetric-key and public-key cryptosystems, data is encrypted such that only authorized users can access it. For performing operations on the encrypted data, data must be decrypted before performing operations. Similarly, to take advantage of the cloud provider's analytic services on data stored on the cloud, whenever an operation is required such as query, the cloud provider needs to decrypt the data first and perform the required operation resulting in familiarizing a third party (cloud provider) with the content of the data (Acar, Aksu, A. Selcuk Uluagac, & Conti, 2016). A Homomorphic Encryption is the conversion of data into ciphertext that can be analyzed and worked within its encrypted form. It enables computing meaningful operations on the encrypted data without observing the actual data (Kocabas & Soyata, 2014). The main goal of the homomorphic encryption is to prevent rogue insiders from violating the privacy of information that is in the hand of a third party such as the cloud provider.

The non-cryptographic approaches mainly use policy-based authorization infrastructure such as access control policies to enforce privacy control to information. There are a number of access control mechanisms that were in the literature, they aim to control access according to specific requirements that vary in different contexts. Examples of access control mechanisms are Role-Based Access Control (RBAC) proposed in (Bertino, Bonatti, & Ferrari, 2001)(Sandhu, Coyne, Feinstein, & Youman, 1996), Attribute-Based Access Control (ABAC) proposed in (Yuan & Tong, 2005), Mandatory Based Access Control (MBAC) proposed in (Hu, Ferraiolo, & Kuhn, 2006), and Identity Based Access Control (IBAC) (Gupta & Quamara, 2018). Each mechanism satisfies the need for controlling the access to information according to different requirements and based on different variables. However, the authors in the survey conducted in (Chenthara, Ahmed, Wang, & Whittaker, 2019), the conclusion reached in the review was that with all the privacy-preserving mechanisms available in the literature, a breakthrough in research to sustain the confidence and credibility of patients is essential for the wide-scale usage and success of the digital health care, the authors stated: “Existing smart health solutions provide a certain level of immunity but not a foolproof mechanism”.

2.4 Summary

The chapter aimed to present a review of efforts that have been found in the literature which researchers have put towards protecting the privacy of healthcare information. The chapter was presented in logical parts; the first part presented the challenges encountered that hinder the current healthcare information system from interoperating with each other to facilitate means of collaboration in terms of collaborative use of patients’ information. Section (2.1) presented the interoperability challenge that is hindering the current healthcare information systems from seamlessly sharing and collaboratively using information. The interoperability challenge was explained and the efforts that researchers have put towards overcoming it were reviewed. The conclusion derived from the section was that the current healthcare information systems lack interoperability due to a number of issues. The solution for enabling the share of healthcare information was seen as possible in a decentralized architectural design rather than a centralized one. Section (2.2) introduced the anonymization challenges encountered when using patients’ information for research purposes. Various privacy-related attacks that are performed on aggregated patients’ information were reviewed in the section. The anonymization models and techniques that are found in the literature were also highlighted and explained in terms of how they work. Privacy attacks on each privacy model have been

explained. The conclusion derived from the section is that none of the privacy models may lead to a sufficient anonymization level of patients' information but at the cost of the utility of the data. They are also prone to many privacy-related attacks that could violate the privacy of individual patients when using aggregated information for research purposes. To date, data anonymization remains a bleeding edge in the area of using patients' information for research purposes in a privacy-preserving manner. Section (2.3) presented various topics related to the technology of cloud computing. Cloud computing is found the dreamed vision of healthcare information systems. The technology of cloud computing was introduced, its deployment models and its immaturity in terms of security and privacy protection were discussed and explained. The section reviewed various privacy-preservation techniques to protect the privacy of information in healthcare cloud applications. The conclusion derived from the section is that the adoption of cloud computing in the healthcare domain has not yet been achieved due to challenges related to the security and the privacy of information.

Chapter 3: Research Methodology and Design

The research aims to produce a cloud architecture design that is less vulnerable to the privacy attacks identified in the literature concerning the utility of the shared information. The research was conducted in a multi-methodological approach underpinned by the Design Science (DS) research methodology described in (Peffers, Tuunamen, & Rothenberger, 2008). The Design-science paradigm seeks to extend the boundaries of human and organizational capabilities by

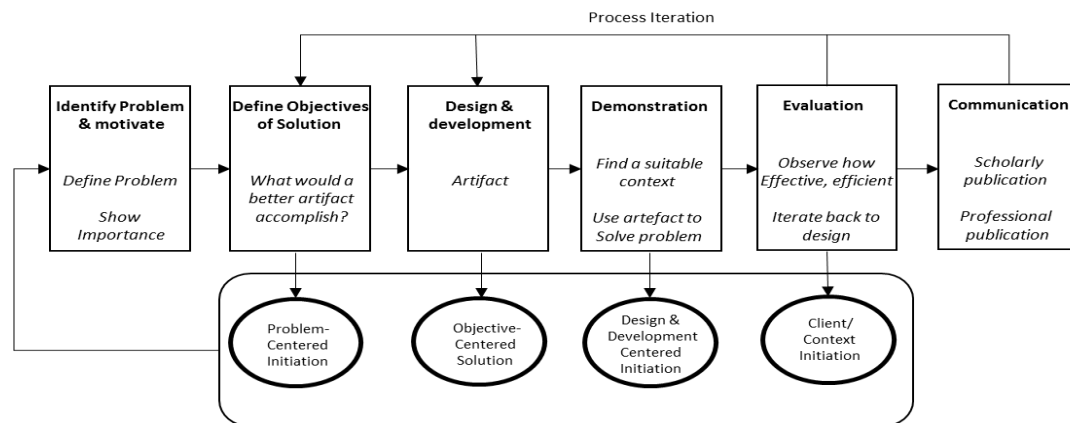


Figure (3.1) Design Science Research Methodology
(Peffers, Tuunamen, & Rothenberger, 2008)

creating new and innovative artefacts (Hevner, March, Park, & Ram, 2004). DS research methodology consists of six main activities in a nominal sequence namely; problem identification and motivation, the definition of the objective for a solution, design and development, demonstration, evaluation, and communication. Figure (2) depicts the sequential activities in the design science research methodology.

The authors in (Peffers, Tuunamen, & Rothenberger, 2008) state that although the DS methodology comprises activities that are structured in a nominal sequential order, it is not expected that researchers would always follow the sequential order from activity 1 through to activity 6. The starting activity can be decided according to the approach of the research. In the Problem-centered initiation, the idea of the research is a result of a problem observation or a suggestion of future work in prior research or project, the researchers therefore might start with activity 1 and proceed in this sequence. In an Objective-centered solution, the idea of the research can be an industrial or research need that can be addressed by the development of an artefact. This can make researchers start with activity 2. In the approach of Design and development-centered, the researchers may start with activity 3. The idea of the research can be a result of an artefact existence that has not yet been formally thought through as a solution

for the explicit problem domain in which it will be used. Such an artefact might have come from another research domain, it might have already been used to solve a different problem, or it might have appeared as an analogical idea. Finally, in the client-/context-initiated solution, the idea of the research is based on observing a practical solution that worked. This results in a DS solution if researchers work backward to apply rigor to the process retroactively. Researchers may opt to start with activity 4. This could be the by-product of consulting experience.

3.1 Research Design

This research aimed to propose a solution to the problem of privacy preservation when sharing patients' information for various purposes such as research and healthcare betterment. This problem has been observed and suggested as future research opportunities in the literature. Therefore, concerning DS research methodology, this research will be conducted following the Problem-Centered Initiation Approach starting from activity 1 through to activity 6 in reference to the CATCH research project (Studnicki, Steverson, Myers, Hevner, & Berndt, 1997). Figure (3.2) depicts the sequential activities of this research design.

While sharing and using healthcare records have a significant potential to facilitate research and improve the quality of medical care provided to patients, privacy is considered a

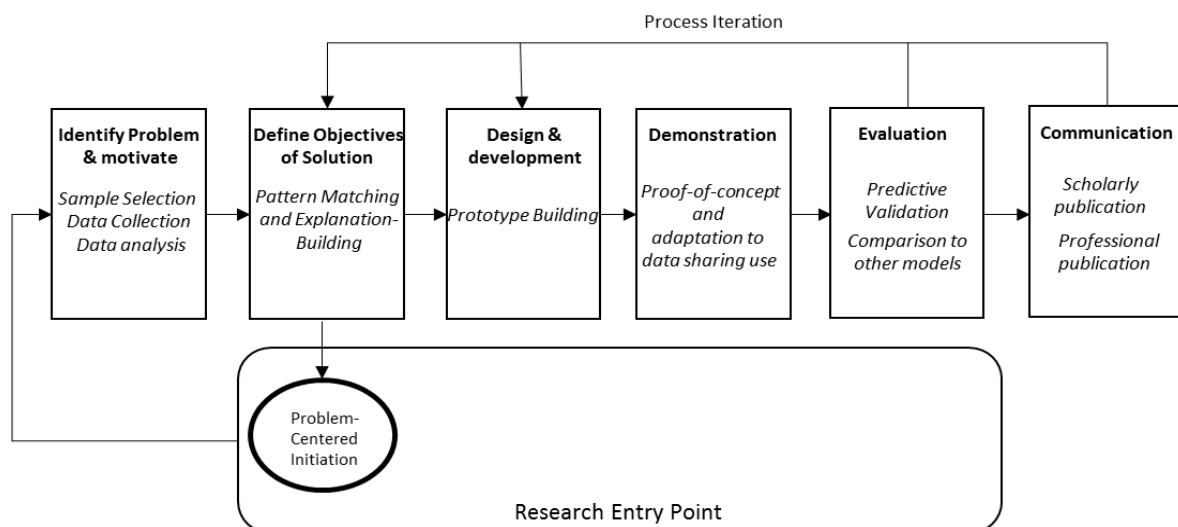


Figure (3.2) Design Science Research Methodology

challenging issue that hinders patients' acceptance to sharing their health-related records. The need for a privacy-decentralized design of architecture to share and use health-related information has triggered the conduction of this research.

Activity 1: Problem Identification and Motivation Activity

The amount of health-care records and health-related datasets is increasing rapidly due to the advanced and relatively cheap data collecting methods such as mobile devices, wearable sensors, and home wireless networks. This data generates special value when it is shared and collaboratively used among different parties involved in the healthcare area. For example, it can facilitate research on rare diseases; assess the current healthcare systems and distribution of resources to better understand the effectiveness of medical treatments and predict various medical conditions. However, sharing such information introduces a risk of patients' privacy breaches, data protection failures, and other related issues. Therefore, many policy-related issues such as privacy policies must be addressed before the full potential of such data can be obtained. For that, this research intends to investigate the current practices followed by healthcare information systems for sharing healthcare information, understand how data can be shared according to its intended utility in a privacy-preserving manner. The research will inquire as to the way healthcare practitioners use shared information for medical care improvement. Following the criterion strategy, a sample of four healthcare-related institutions will be selected to study. Many medical practitioners from different healthcare-related institutions were invited to participate in the research. The criterion used for the selection required each selected institution to have a need to use and share healthcare information. The data was collected through literature review and open-ended interviews with research participants. The collected data was systematically filed and prepared for the analysis phase. The data were analyzed within the context of each case and patterns are to be extracted.

Activity 2: Defining the Objectives of the Solution

The main objective of this research was to propose a cloud-based architecture design for sharing healthcare information in a privacy-preserving manner. The major challenge was twofold; preserving the privacy of patients' information while facilitating the sharing of such information among parties involved in healthcare. The solution was proposed as a cloud-based architectural model that can be employed in healthcare information systems. The model aimed to overcome the privacy challenge when sharing patients' health-related information as a contribution to obtain the benefits of sharing patient's health information such as medical care betterment and research. For that, an explanation-building technique was employed to match the patterns extracted in the previous research activity to the theoretical proposition of the study. The results of the pattern-matching indicated the characteristics of the intended model.

The resulting characteristics will draw the definition of the solution that was considered in the model design and development activity.

Activity 3: Design and Development Activity

The desired artefact in this research was a cloud-based architectural model that enables the adoption of cloud computing technology in the healthcare sector to share information and achieve sufficient collaboration among the involved parties. The search for an effective artefact requires utilizing available means to reach desired ends while satisfying laws in the problem environment (Hevner, March, Park, & Ram, 2004). The researcher attempted to draw from the current cloud-compatible privacy protection techniques proposed in the literature concerning the need for sharing the data. The characteristics identified in the previous research activity was considered in the design of the intended model. The design included two levels of granularity: the privacy-related characteristics of the model and the satisfaction to the need of sharing the information. The design of the model was continuously refined until the model objective was successfully achieved.

Activity 4: Demonstration Activity

After the development of the proof-of-concept-level prototypes; the artefact will be extensively adapted to data sharing use. In (Nunamaker, Chen, & Purdin, 1991), the authors state that when the proposed solution of the research problem cannot be proven mathematically and tested empirically, or if it proposes a new way of doing things, researchers may elect to develop a system to demonstrate the validity of the solution, based on the suggested new methods, techniques, or design. For this research, the researcher will attempt to use a randomly generated healthcare dataset to share using the developed model. Various scenarios of sharing health-related information will be deployed to test the model's ability to share information in a privacy-preserving manner. The demonstration focused on two aspects (1) ability of the proposed architectural design to share healthcare information without questioning the privacy of this information; (2) compliance with the information privacy-related regulations when using patients' information for both; medical treatment purposes and research purposes.

Activity 5: Evaluation Activity

The evaluation activity was conducted based on the model's validity in terms of sharing healthcare data in a privacy-preserving manner. The validity of the model referred to the substantiation that the model, within its applicability in sharing healthcare data in a privacy-

preserving manner, possesses a satisfactory range of accuracy consistent with the intended application of it. Therefore, the validity of the model was tested based on two main facets namely; ability to store patients' information and facilitate collaborative use of it for genuine reasons, compliance with information privacy regulations, and users' satisfaction in terms of the usefulness of the shared data.

In (Sargent, 2009), the author has listed and described a number of model validation techniques that can be adopted individually or in combinations to validate models, they are Animation, Comparison to Other Models, Degenerate Tests; Event Validity, Extreme Condition Tests, Face Validity, Historical Data Validation, Internal Variability, Multistage Validity, Operational Graphics, Parameter Variability, Predictive Validation, Traces, and Turing Tests. However, for the purpose of this research, and due to the nature of the problem under study, the Predictive Validation was the technique adopted for the validation of the intended model.

In the predictive validation technique, the model is used to predict the system's behaviors and then comparisons are made between the system's behavior and the model's prediction to determine if they are the same. This technique was used to compare the output of the system with the newly designed model against the expectations of it. The intended cloud-based architectural model was expected to facilitate sharing health information with compliance to privacy-related regulations such as the Information Privacy Act.

Activity 6: Communication

The final outcome of this research (Thesis) will be kept at the AUT Library. It will also be published in academic journals, academic conference proceedings, and professional outlets.

3.2 Case Study Approach

This research involved investigating several cases in which sharing healthcare information is needed, and privacy is vulnerable to attacks. Several medical practitioners from each selected case were invited to participate in the research. To gain more profound insights from the case study participants, the primary data in this research were collected through open-ended face-to-face interviews. From the received data, case-based assertions were made and compared in cross-analysis of the studied cases. An inductive qualitative analysis described in (Thomas, 2011) was used to identify findings.

Several well-known case study researchers such as Robert K. Yin and Robert E. Stake have written extensively about case study research and suggested techniques for organizing and conducting such research successfully. However, for the purpose of this research, the research design included the five components of effective case study research specified in (Yin, 2009) namely; research questions, the propositions or the purpose of the study; unit of analysis; the logic that links data to propositions; and finally, the criteria for interpreting the findings.

3.2.1 Research Questions

In this research, the primary intention was to produce a data-sharing model that preserves the privacy of individuals concerning the utility of the shared dataset. This research aimed to answer three central questions namely, “How can patients be sure that their information privacy is protected? and what information to reveal for statistical analysis by cloud providers?” And “How do we maintain the privacy requirements of healthcare data while it is stored in the cloud?”

3.2.2 Research Proposition

Due to the explanatory nature of this research, research propositions are of a vital necessity; the research questions needed to be translated into proposition as suggested in (Rowley, 2002). Translating the research questions to propositions has helped the researcher to structure the data collected and analyzed to satisfy the purpose of the research. The goal of this case study research was to derive and understand the characteristics required for the data-sharing model that satisfies the need of it for medical practitioners and statisticians with a guarantee to preserve the privacy of patients. The researcher attempted to make speculation based on the literature as to what the findings were expected to be.

3.2.3 Unit of Analysis

The multiple case study methodology is believed to be more suitable for studying typical cases of information systems implementations (Shakir, 2002). The unit of analysis as described in (Yin, 2009) is the area of focus that a case study analysis. For this research, the unit of analysis is the practices used in healthcare information systems to access patients’ information and preserve their privacy when sharing their health-related information. This unit of analysis is directly tied to the main research questions. Yin also wrote that an appropriate unit of analysis occurs when primary research is accurately specified.

The selected cases for this research were analyzed for the goal of generalizing the sufficient characteristics of privacy-preserving models for sharing healthcare information without affecting the accuracy of the shared information. The nature of this research relies on analytical generalization, which Yin (1994) has defined it as the generalization of a particular set of results to some broader theory. Yin states that the selection of multiple case studies when the generalization is to happen analytically needs to follow the replication logic.

For applying the replication logic, two approaches can be followed according to the context of the research namely; literal replication and theoretical replication (Cavaye, 1996) (Yin, 1994). In literal replication, the chosen cases should have similar settings and are expected to achieve similar results, while theoretical replication requires selecting cases that have different settings and are expected to produce different results. For this research, the literal replication approach was used because cases chosen will have similar settings in terms of sharing patients' information, and the privacy of patients is expected to be violated in all cases.

The requirements of the replication logic for multiple-case design also provide suggestions for deciding the number of cases to be studied. The satisfactory number of cases in a theoretical replication is six to eight cases and three to four cases for literal replication (Yin, 1994). Thus, the sample of cases for this research will comprise 4 cases of healthcare-related institutions to be studied.

However, the replication logic alone cannot methodologically guide the process of selecting cases for the research. There is a need for sampling strategies to be followed to have an accurate case selection process. In (Patton, 1990), the author introduced sixteen sampling strategies which can aid the process of selecting cases namely; extreme case, intensity case, maximum variation, homogeneous, typical case, stratified purposeful case, critical case, snowball, criterion, theoretical, confirming and disconfirming, opportunistic, random purposeful, politically important case, convenience, and a combination strategy.

The quality of the multiple-case research design relies heavily on the process of selecting case studies. For the research quality regarding the design, the process of selecting case studies should be driven by the two issues; appropriateness and adequacy (Kuzel, 1999). Appropriateness is highly related to the demonstration of a fit to the purpose of research and the phenomenon of inquiry, while adequacy is about how much is enough or how many cases to be studied (Patton, 1990) (Kuzel, 1999). For this research, the criterion sampling strategy is followed for selecting the cases. In criterion strategy, cases were decided if they were

information-rich and might reveal a major system weakness that could be improved (Patton, 2001).

3.2.4 The logic that links data to propositions

The logical linkage was made after the data collection phase of the research as themes emerged. The data collected for this research included a multitude of different pieces of evidence from different sources. The data collection phase of this research is conducted through a literature review and intensive open-ended interviews with research participants. Interviews are one of the most important sources of case study information that serve to validate previously collected or available data (Tellis, 1997). Open-ended interviews can offer richer and more extensive materials than data from surveys or even the open-ended portions of survey instruments (Yin, 2009). The data collected from the interviews were systematically filed to prepare it for the analysis phase as suggested by (Berg, 2004). The data were analyzed within the context of each case.

In this research, the analysis of the data relied on the theoretical propositions that led to the case study. The explanation-building analysis technique presented in (Yin, 1994) was employed in this research to establish the quality of the research. It is a form of pattern-matching, in which the analysis of the case study is carried out by building an explanation of the case (Tellis, 1997). Nevertheless, because there are no standard procedures for the analysis of the case study results, the analysis of the case study for this research adhered to the principles outlined in (Rowley, 2002) for a good case study analysis which state that the analysis should: make use of all of the relevant evidence, consider all of the major rival interpretations to explore each of them in turn, address the most significant aspect of the case study, draw on the researchers prior expert knowledge in the area of the case study but in an unbiased and objective manner. After the analysis of the collected data, the researcher attempted to match patterns extracted from the analyzed data to the theoretical proposition of the case study.

3.2.5 Criteria for interpreting findings

The last component of an effective case study research design is the criteria for interpreting results. In Case Study research, the coding of the data should happen before the themes development (Yin, 2009). Therefore, data codes were generated before the data collection and analysis phases of the research. Once themes emerged, the researcher extracted meaning from the findings and determined the characteristics of the intended privacy-preserving data sharing

model concerning the legal frameworks such as the Health Insurance Portability and Accountability Act (HIPPA) and Data Protection Act in New Zealand.

3.3 Data Analysis

Stake reminds qualitative researchers that there is no particular moment when data analysis begins. He defines analysis as the process of deconstruction of data and impressions. It then entails giving meaning to the parts (Stake R. , 1995). Qualitative research studies involve a continuous relationship between data collection and data analysis (Corbin & Strauss, 1994). For this reason, the researcher in this research began the analysis of the data directly after the first interview. This has helped to identify patterns and facilitate subsequent data collection as recommended in (Corbin & Strauss, 1998).

The analysis of data in qualitative research is seen as the process that involves preparing and organizing the data and reducing them through extensive coding (Creswell, 2007). For this research, due to the iterative nature of the data analysis, the spiral is used as a metaphor to illustrate the analytic process. The spiral model starts with organizing the data, continues with reading and writing notes, and moves on to describing, classifying, and interpreting. The spiral progressively thickens to reflect how through repeated reading, interpretation, and coding, a new understanding of the data is achieved. This provides the foundation for the development that follows and positioning of the new theory against the present research (Gregor, 2006).

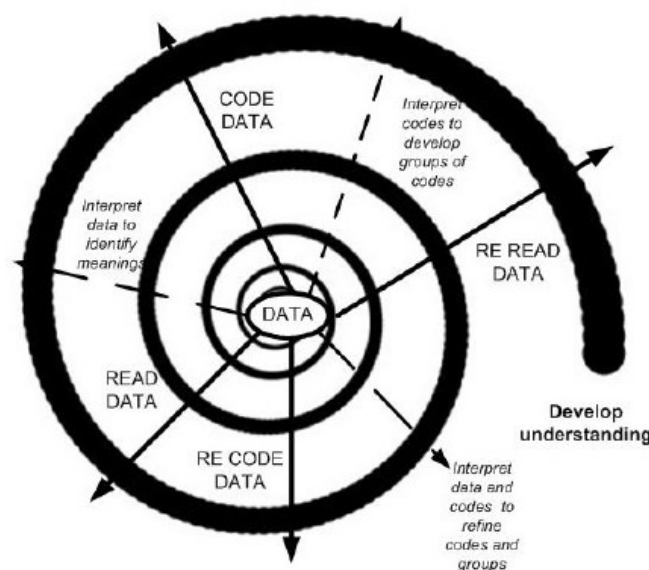


Figure (3.3) Data Analysis Model (Creswell J. , 2007)

After each interview, the researcher created documents from the notes taken during and after each interview. Each interview was transcribed into Word document and kept together with the notes on a portable computer and backed up on different locations (USB) drive. A password was set for all the documents to assure the confidentiality of the interview transcriptions.

The analysis of the data followed Creswell and Yin and Stake's model of data analysis; direct interpretation of the data was conducted as well as aggregation of instances in the form of codes. Stake suggests that some issues require categorical analysis, while other issues may occur once and require direct interpretation (Stake R. , 1995).

To best address the research questions, the analysis was conducted overall interview transcripts. Each transcript was firstly described but not analyzed as suggested by Yin (2003). Initially, the researcher conducted a preliminary exploratory analysis with all interview transcripts and notes. During this initial stage of analysis, all transcripts were read through and notes were made. And to better understand the transcripts from the participants' perspectives, during this phase of the analysis, the research questions were set aside (Creswell, 2007).

The analysis of data for this study followed the Creswell's hierarchical approach (Creswell J. W., 2009) which involves six steps process building from bottom to the top. Although the steps are described in linear order, Creswell described it as an interactive in practice, meaning that the process is not static. He explains "the various stages are interrelated and not always visited in the order presented".

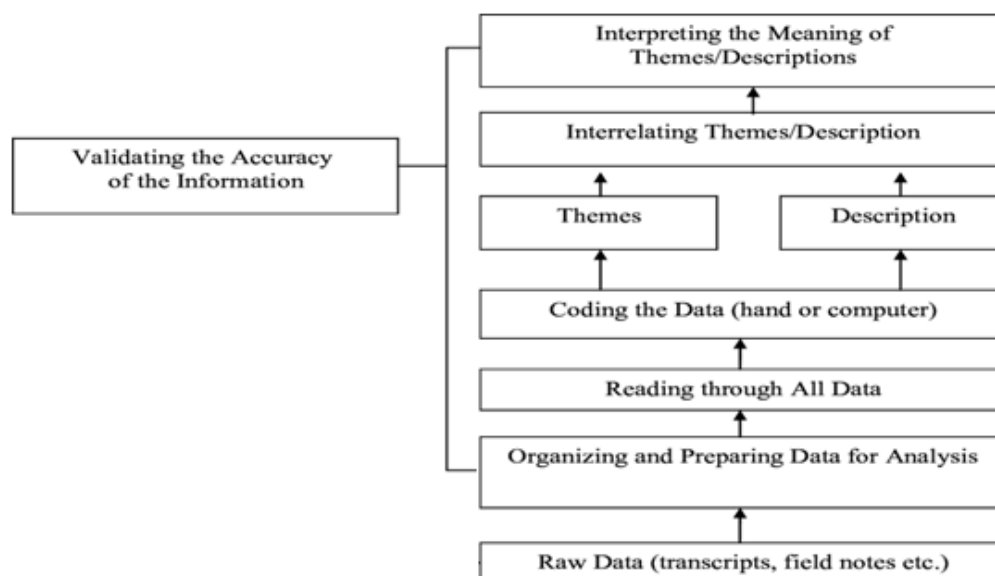


Figure (3.4) Creswell's hierarchical approach (Creswell J. W., 2009)

Step 1. Organize and prepare the data for analysis. During this step, audiotapes from interviews were reviewed. Interviews were all transcribed into Word documents following an intelligent verbatim approach.

Step 2. Reading through all data. This step aims to provide a general sense of the information in hand. In this step, the researcher reflected on the overall meaning of the interview transcripts. As mentioned earlier, the researcher has read through each interview transcript from the beginning to the end a few times without considering the research questions to understand the interview transcripts from the participants' perspectives. The researcher aimed to immerse himself with the data in hand. This aligns with Esterberg's directive statement "get to know your data"(Esterberg, 2002).

Step 3. Coding the data (hand or computer). The authors in (Rossman & Rallis, 2016) define coding as the process of organizing the data by bracketing chunks and writing a word representing a category in the margins. Esterberg (2002) says "in qualitative analysis, the goal is not to assign numbers to a case, rather the goal is to begin to focus on the potential meanings of your data" and he suggests that coding is the first step in making sense of the data. In qualitative research, the development of interpretation happens through coding. In this step, the researcher followed Creswell's procedure of organizing the materials into segments by taking the text data and segmenting sentences into categories. For this, the eight steps coding process proposed by Tesch (1990) was followed.

1. *Get a sense of the whole.*

The researcher has read all the transcriptions carefully and ideas formed. Each transcript was read twice for the goal of getting ideas and understand what categories to extract.

2. *Pick on the document (transcript) and go through it asking yourself, "What is this about?"*.

The researcher in this step read different transcripts and wrote his thoughts about them in the margin. Transcripts were organized by grouping responses according to the interview questions.

3. *When you have completed this task for several participants, make a list of all topics, cluster together similar topics.*

As mentioned earlier, participants' responses in all transcripts were grouped according to the interview question number. After gaining familiarity with the transcripts and searching for

ideas, coding was firstly done with an eye for both descriptive and thematic data as suggested by Creswell (1995). Next, responses for all questions were examined deductively to organize them in categories. The material was organized by segmenting sentences/paragraphs into categories. 10 categories were derived namely; Information required in emergency cases, information required in an out-patient clinic, information that requires permanent storage, information that does not require permanent storage, information that can be deleted to anonymize records, most private information, parties require accessing the most private information, information that is promptly needed in emergency cases, problems with the current system, desired characteristics.

4. Now take the list and go back to your data. Abbreviate the topics as codes and write the codes next to the appropriate segments of the text.

Three rounds of interpretation and coding were conducted. The data were inductively and iteratively coded to achieve a deeper understanding of the data in each category formed in the previous step. Data of related meaning were grouped for defining data codes.

5. Find the most descriptive wording for your topics and turn them into categories. Look for ways of reducing your total list of categories by grouping topics that relate to each other.

The 10 categories and data represented in each of them were further examined in the hope to find common themes across the categories. The categories were grouped into a more comprehensive set of themes that represented the key points made by research participants. The themes that emerged are information storage, information disclosure, information accessibility, information significance for research.

6. Make a final decision on the abbreviation for each category and alphabetize these codes

The final decision on the abbreviation for each theme was made. For this research, the main purpose of the data collection and analysis was to identify the required characteristics of the intended data-sharing model. Therefore, the naming of themes was made technical to suit the purpose of the study and better serve in later phases of the research which included the data-sharing model.

7. *Assemble the data material belonging to each category in one place and perform a preliminary analysis.*

The data were fully separated and organized under its corresponding themes. Each theme and its associated data were put separately in one place to undergo intensive analysis. Separate documents for data themes were generated as preparation for analysis. The analysis of data in each theme aimed to extract the required characteristics of the intended model.

8. *If necessary, recode your existing data.*

The coding process was iterative. It was necessary to recode the data a few times until the current codes were achieved. As the researcher has been through the data, codes often were not accurately grouping data, many rounds of analysis were conducted until the final codes were achieved.

Step 4. Using the coding process to generate a description of the setting or people as well as categories for analysis. In this step, the researcher used the coding process to generate descriptions for the codes and its associated data. The descriptions of the codes led to generalizing several themes that emerged. The emerged themes were then analyzed, and more generalized descriptions were derived.

Step 5. Advance how the description and themes will be represented in the qualitative narrative. In this step, the emergent themes were woven narratively, and findings were extracted logically from the participants' responses in interviews.

Step 6. Interpretation of data meaning. Creswell recognizes that the researchers' experience and background knowledge play a significant role in the meaning-making process. Having conducted previous studies on healthcare data sharing, the researchers' background was enriched regarding what data is shared and why it is needed. The interpretation of the data meaning was supported in the literature. The information gleaned from the literature as compared to the findings of the data interpretation.

3.4 Summary

This chapter aimed to present the methodological approach followed for the conduction of this research. The research design was first outlined with justification to how each phase of the research has contributed toward achieving the outcome of the research. Section (3.1) presented the design of the research in terms of its sequential activities, each activity was explained and

the goal of it was outlined. Section (3.2) presented the case study approach which was part of the research design. The section aimed to explain how the case study research approach contributed to understanding how patients' information is being used for medical treatment and research purposes. The main purpose of the case study approach was to understand how information systems can best serve the healthcare domain in terms of sharing information. Each component of the case study research was explained and justified in terms of its applicability to the current research. Finally, section (3.4) presented the process followed for analyzing the data in this research. The chapter presents information about the data gathering protocol, analysis discussion, and findings.

Chapter 4: Data Collection and Findings

Different studies have indicated the influence of information sources to know the patient or know the medical condition of a particular patient in healthcare settings (Blythe & Royle, 1993) (Zhou, Ackerman, & Zheng, 2009). The availability of information in any healthcare setting plays a significant role in the quality of care provided to patients. A comprehensive, methodical review of pre-procedural care and management in patients undergoing vascular and interventional radiology procedures is presented in (Taslakian, Sebaaly, & Al-Kutoubi, 2016). The authors state that reviewing appropriate diagnostic tests, imaging studies, and medical history ensures that the proper procedure is selected and is indicated when providing health-care to any patient.

This chapter presents a case study research approach that was conducted to understand the desired characteristics in healthcare information systems. The main goal was to understand how information systems in the healthcare domain best serve towards the improvements of healthcare services provided to patients. The identified characteristics in this study were appreciated in the design of the cloud architecture for storing and sharing information related to patients and their health.

4.1 Data Gathering

This section presents the study sample and the process of the interview. It also identifies the sample characteristics in light of the research methodology and model.

It was needed to build participant samples that can provide meaningful and relevant information. The research participants were recruited from amongst employees of organizations involved in the healthcare sector and needed to deal with patients' information in their daily business operations. A total of 18 organizations were approached, of which 6 of them accepted to participate in the research. As a result, 19 individuals from different organizations were recruited for the research. Despite the difficulties in gaining access to research participants, the recommended sample size of 4 to 6 organizations was met. Table 4.1 presents a list of the participating organizations with their identifiers which will be used further in the text as references to each participant.

Organization Code	Organization business Description	Number of participants interviewed
ORG1	Tertiary healthcare institution	4
ORG2	Hospice hospital	3
ORG3	General practice and urgent care institution	4
ORG4	General practice and urgent care institution	5
ORG5	Urgent and Ambulance care	1
ORG6	Pharmacy	1

Table (4.1) Participating Organizations

During the data collection phase of this research, participants were recruited following the procedure outlined in the application for ethical approval to the Auckland University of Technology Ethical Committee (AUTEC), see (Appendix B). The process of collecting the data was conducted following the guidelines for conducting ethical qualitative research described in (Schutt, 2009) which are: the participation in the research is voluntary, participants are well informed about the research, participants are invited to sign a consent form prior to the interview, identification details are not recorded in the interview transcript to protect participants' privacy, participants have the right to decline to answer any question. A copy of the ethics approval from the AUTEC along with the participant information sheet was made available to each participant before the interview. The research participants were interviewed in locations according to their convenience. Most of them were interviewed in their offices during their working hours. Two participants preferred interviewing outside their working hours and at the alternative district office site. All interviews were conducted in a face-to-face manner and lasted from 25 to 45 minutes.

As the first step of each interview, the researcher introduced himself to participants to establish rapport and gain participants' trust as recommended (Seidman, 2006). Before starting each interview, an overview of the questions was briefly presented to the participants as recommended by (Patton M. Q., 2002) as some questions contained more than one idea. A hard copy of the interview questions was made available to each participant during the interview and was used as a guide throughout the interview time.

With the participants' approval, interviews were all audio recorded to ensure accurate transcription. Recording the interview makes it easier for the researcher to focus on the interview content and verbal prompts (Jamshed, 2014). Recording the interviews enabled the researcher to generate a verbatim transcript of each interview. The researcher also took brief

handwritten notes during each interview for tracking key points and highlighting ideas that have importance in the context of the research.

4.1.1 Interview Protocol

The research collected data from research participants using standardized open-ended interview protocol. All interviewees were asked the same basic questions in the same order and questions were worded in a completely open-ended format. This method was seen fit for this research due to a number of reasons explained in (Patton M. Q., 2002) which are: Respondents answer the same question which increases the comparability of responses, data are complete for each person on the topics addressed in the interview, the interviewer effects and bias when is reduced especially when several interviewers are used, allows evaluation users to see and review the instrumentation used in the evaluation, and finally facilitates organization and analysis of the collected data. Moreover, this method was used in the research because it allowed efficient use of participants' time as getting access to practitioners was constrained due to their busy working time.

However, the author in (Patton M. Q., 2002) suggests that this method may constrain the naturalness of the questions and the answers; this limitation was eliminated in this research by asking participants to elaborate more on points they made during the interview.

4.1.2 Data Transcribing and Preparation

Before analyzing the data, each interview recording was transcribed into documents using MS Word. The intelligent verbatim approach was followed for the transcribing interviews in which the richness of the responses was preserved and the meaningless utterances such as “ahh” or “hmm” were removed. As recommended in (McLellan, MacQueen, & Neidig, 2003), the researcher has verified the accuracy of the transcription three times for each interview by comparing each transcription to its original source (recording). Each transcript included the question numbers and their corresponding answers. Questions were numbered in the same sequence in all transcripts. To assure the confidentiality of research participants, the organizations' names and participants' names were removed from the transcripts.

4.1.3 Organizing the Data

Firstly, the researcher explored the data collected in the interview to check the expected mapping of answers to the questions asked in the interview onto the perspectives of the research

framework and the research questions RQ1, RQ2, and RQ3. Reading the data confirmed that normally an answer to an interview question could contain information that may relate to more than one research question and perspective. The table below illustrates how responses to the questions in the interviews were related to research questions/perspectives.

Research Perspective	Interview Questions	Related Content
RQ1	1, 2, 3, 4, 5	Information storage Information disclosure
RQ2	1,2,3,5,6,7	Information Accessibility Information categories
RQ3	2, 3, 4	Information of significance for research purposes

Table (4.2) Interview questions and their relations to main research questions

Since the gathered dataset was reasonably large, the researcher attempted to organize it in a more manageable manner before commencing systematic analyses and interpretation. For that, Computer-Aided Qualitative Data Analysis Software (CAQDAS) was employed. The use of CAQDAS in this study allowed the researcher to take advantage of data documenting, organizing, and visualizing the capabilities of CAQDAS. The CAQDAS tool chosen for this research was NVivo12 which was supported by the researcher's university Auckland University of Technology.

The interview responses were grouped according to the interview question number. Each response was tagged with the participant's identifier. All questions were answered by participants, some participants provided short answers while others gave long ones.

Choosing a CAQDAS package for any research requires considering the tool's suitability for the research approach, the methodology of the study, as well as the researchers' proficiency in using it (John & Johnson, 2004). The choice of NVivo as a CAQDAS tool for this research was supported in the literature. The authors in (Bazeley & Jackson, 2013) have demonstrated how NVivo can be used in qualitative research to manage data, query data, manage ideas, visualize data, and report from data. NVivo is recommended to use for inductive coding due to its ability to separate source data (transcripts) and codes. It facilitates an efficient iterative

process by allowing multiple updates of the codes and their descriptions, preparing the data for the subsequent theme identification (Frost, 2008).

The researcher had received initial training in NVivo by experts from Academic consulting and had arranged with them for ongoing support throughout the analysis journey of the research. In further sections a detailed account of the research process that shows how NVivo was used to code and analyze data, keep track of the intermediate steps of the coding and analysis, and document and visualize the analysis outcomes.

4.2 Discussion

The main purpose of the research was to propose a cloud-based architecture for storing, sharing, and using patients' records in the healthcare domain in a privacy-preserving manner. The following research questions informed this study: (a) how do we maintain the privacy requirements of healthcare data while it is stored on the cloud? (b) What are the characteristics of a privacy-preserving cloud-based architecture for sharing healthcare information? And (c) what information can be disclosed for statistical analysis by cloud providers?

During the in-depth interviews, the research participants described the need to use patients' health in their daily practices in providing healthcare to patients. They also discussed the types of medical records in terms of their significance and need when providing healthcare to patients. The research findings presented in this chapter are based on the data collected in open-ended interviews with research participants.

4.2.1 Background

The participants in this study were comprised of 19 healthcare practitioners from 6 different sectors in the healthcare domain. All participants required using and sharing patients' records in their daily practices during providing health-related care to patients. They all had experience in sharing and needing to share and use patients' health records. As mentioned earlier, the data was collected via open-ended interviews. Research participants were asked questions regarding the need for patients' information in different healthcare-related settings such as emergency, out-patient clinical visits, and research. They were also asked about the types of patients' information in terms of privacy levels and significance on the accuracy of care delivery. Participants were also given a chance -by responding to a question- in the interview to criticize the current systems they used for sharing and accessing patients' information, the main goal

was to relate their responses to the previous interview questions and better understand and justify the desired characteristics of healthcare information systems.

4.2.2 Interview questions

Question 1: *What information do you as a healthcare practitioner require in order to provide accurate care to a particular patient in cases of emergency? How does it differ from the information that a practitioner in an outpatient clinic requires?*

The main purpose of this question was to gain an overall sense of how patients' information is used by medical practitioners in different cases such as emergency cases. The question had two parts, the first part was objective and limited to emergency cases, while the second part was broader and responds to it were expected to inform the research about how such information is needed in other cases. Out-patient clinical visits vary in purpose therefore, the question was expected to bring out indicative information about how patients' medical records can be helpful in different cases.

Question 2: *What type of information that is required to store permanently in any patient's health record? For example, if a patient has suffered from a heart attack at a certain time of his/her life, how important is it to store information about such incident in the record of the patient, and when is it needed?*

The purpose of this question was to gain knowledge about the information that has an ongoing impact on patients' health assessment and care. The intention was to achieve a closer view of information that is considered critical in terms of their impact on the quality of medical care provided to a patient. The responses to this question were expected to feed the research with information that helps to identify an information category in a patient's medical record.

Question 3: *What type of information does not require being stored permanently in any particular patient's health record? And if there is a need for temporary storage, for how long such information is required to be stored for future access to assure accurate health care for the same patient?*

The purpose of this question was to gain knowledge about the information that does not have an ongoing impact on patients' health assessment and care. Like the previous interview question, the responses to this question were expected to feed the research with information that helps to identify an information category in a patient's medical record. The intention was

to achieve a closer view of information that is not considered critical in terms of their impact on the quality of medical care provided to a patient.

Question 4: *If researchers or data analysts are to have the information in anonymized form for research purposes, what particular data fields (e.g. Birth Date) can be removed without affecting the outcome of the research?*

The main purpose of this question was to identify an information category that can be deleted from any patients' health records without affecting the quality of the research. Some information in each patient record is used for describing the patient, patient's personal preferences, and facts about the patient; such information does not have any need when assessing the patient health. This question was asked in the hope to derive more knowledge about this information and help in designing how patients' records can be disclosed in an anonymized manner without affecting its usability in terms of research.

Question 5: *Some information is considered the most private information that a patient would not want to disclose such as information relating to mental health, sexual health, and alcohol/drug addiction. Who would be most interested in accessing such information and in what cases?*

This question aimed to identify the information that a patient would be most concerned about its disclosure. The question was worded to ask about the most private information and parties that require accessing it, however, the purpose in this question was twofold; first to identify the information and the parties who require accessing it, and secondly to know the cases in which accessing such information was not required.

Question 6: *What information is promptly needed to assist a patient in an emergency instance? Is there information related to patients' health that is required to be accessible promptly for every patient's visit? What is it?*

The purpose of this question was to identify the information that was always needed regardless of the purpose of the patient's visit. For example, patients sometimes visit medical institutions for a regular check-up or checking for the existence of certain diseases such as breast cancer. Such visits may not require accessing all information about the patient's health. The question intended to derive an information category that was needed to be accessible for every patient's visit. The question was worded to include the repetition of question 1 in the first part of it to

assure that participants understand and distinguish between the emergency case and the information required for every patient visit.

Question 7: To effectively get the benefits of information systems in the healthcare domain, what are the characteristics required for an information system in the healthcare domain to improve the quality of healthcare provided to patients?

The main purpose of this question was to allow participants to criticize the systems that they were using in their daily healthcare practices. The intention was to understand and identify the desired characteristics of healthcare information systems to be considered in the intended cloud architecture design. The responses to this question were expected to highlight the downfalls of the current systems from medical practitioners' perspectives to address them in the intended design of the cloud architecture.

4.3 Study Findings

Two main themes emerged from the data analyzed namely, Information needs and desired system characteristics. While these themes are reported as being discrete, there was considerable overlap among them. The participants' responses to interview questions were often addressing more than one theme. The description of data was made as to where they appeared to be most logically fit, see (Appendix C).

4.3.1 Information Needs

Each of the research participants spoke of ways in which they needed to use and/or share patients' health-related information based on the duties they had in terms of providing healthcare to patients. No participant questioned the benefits of sharing patients' health-related records in providing healthcare to patients. They all explained how information could help them provide accurate healthcare to patients. Some also explained how lack of information could cause serious problems to patients, delays in providing medication and taking right actions, delays in understanding conditions that patients presented with, and often giving the wrong medication to patients. Participants explained that all information related to patients' health including diagnosis and medication is important, however, some information related to minor incidents such as minor injuries may often have no significance on the patient's health in the future and neither needed for potential healthcare for the patient. Moreover, despite the importance of all information related to each patient's health conditions and assessment records, the analysis of the data allowed to identify four main categories of patients' health

information in terms of their need when providing healthcare to patients. These categories are Information that is constantly required in every patients' visit, information that is required in patients' emergency visits, information that is required in out-patient clinical visits, and information for research purposes.

a) Information that is constantly required in every patients' visit (All_V)

Every participant interviewed responded to interview questions as per their knowledge in the medical field, personal experience, and professional experience in providing medical care to patients. By referring to the interview questions, questions 1 and 6 asked objectively about the information that a medical practitioner needed when assessing a patient for different purposes such as in emergency settings, out-patient clinics, and other visits.

Few terms were repeated among most participants' responses such as patients' identifications and demographics details such as age and sex, current medications, significant medical history, and conditions. Some participants explained the need for this information and how important to have it handy in every patients' visit. Regarding patient identification, participant 15 said: *"So for every patient's health visit you want to confirm that this is the right patient you are seeing so you need to have the patient's name, date of birth, their address, NHI number particularly if we're ordering any investigations because that is what's needed to be put on any like blood tests or ultrasound and things we need that information on there. So that needs to be accessible with every patient visit"*. This can also be illustrated by the following quote from participant 1 *"The identification of the patient is the most important, it's more important than any other piece of information"*.

Current medication is another piece of information that was identified as needed for every patients' visit. Almost all participants stated that current medication is important information to know for assessing or prescribing medicine to any patient. The importance of knowing about the medication that a patient is on lies because all drugs come with side effects of which can lead to patient complaints. Moreover, active ingredients in the various preparations can clash causing side effects.

Significant medical history and conditions is another piece of information that all participants said it was required for every patient visit. Participants explained that some past medical conditions might have an impact on long-term care, it also can potentially indicate the overall plan for any patients' care. Participant 15 said *"So that would mean that is kind of like a significant event that has happened in this person's life, and they're going to need some on-*

going treatment for it so, all the medications and the care from now on would focus on preventing something like that happen again. And an event like that his ongoing care that's needed". To further explain the need of knowing significant medical history and conditions for every patient visit, the same participant gave an example incident and said, *"I would say something like a heart attack is a very important thing to permanently have in a patient's record because it has implications for the ongoing management"*.

b) Information that is required in patients' emergency visits (Em_V)

Participants from all the cases clearly stressed the need for accessing patients' information when providing healthcare for any patient in emergency incidents. Most participants in their responses explained how having the right information can significantly help to provide accurate care to patients. This is illustrated by the following statement made by participant 15 *"So when we think about caring for a patient in an emergency, this situation we want to kind of stabilize the patient as quickly as possible, and so there are some crucial aspects I guess that we need to know"*.

It was clear from responses that emergency cases vary from a patient to another, and each emergency incident may require slightly different information to stabilize the patient or identify the symptoms presented. Participant 1 illustrated on this and said, *"I cannot give specific details on which information that I would need but I would look up all of the past clinic notes plus medical record, any previous visits to the hospital because I am in the tertiary sector, radiology laboratory results, discharge summaries allied health."* She also said *"Blood test results such as blood type and that sort of thing as the patient were, for example, needing a blood transfusion or they would look up the past one, they have to have current blood draw for a cross match. So, a previous medical record wouldn't be essential in that case. It's very dependent on the type of emergency that the patient presents with, and it also depends on the conscious level of the patient; if the patient is unconscious but unidentified an electronic medical record will not be of help"*.

Several terms were repeated by all participants when asked about the information required in emergency cases. Along with the information that is required in every patient record (All_V) explained in the previous point, there is other information that appeared to have significant importance when providing care to patients in emergency incidents, this information includes drug allergies, discharge summaries, laboratory results, next of kin details, and blood type. Significant past medical history was found to have significant importance when providing

health care to patients in an emergency case because medical history can have implications for the emergency presentation. This is illustrated in the response of participant 15: *“And any significant past medical history can have implications for the emergency presentation but to stabilize them you don't always have to have the information”*.

c) Information that is required in out-patient clinical visits (OutP_V)

Question 1 in the interview aimed to gain an overall sense of how patients' health information may be needed by medical practitioners in different cases. The question asked about the difference between the information needed in emergency cases and the information needed in out-patients visits. No specific answers about the need for information in out-patient clinics because visits vary in purpose from a patient's visit to another. However, it was derived from participants' responses that the information required in out-patient clinics may differ in comprehensiveness from the information required in emergency settings because patients visiting outpatient clinics are in conscious physical status and stable health conditions in comparison to those in emergency instances. In an emergency, information is needed for stabilizing the patient and take immediate decisions or actions, while in out-patient clinics no instant actions required. This is illustrated in the response of participant 15 *“Apart from kind of the circumstances of the emergency and how it happened the stuff is great from the outpatient clinic because in an outpatient clinic and seeing patients who are stable, they're not kind of needing emergency treatment. And so usually what they are coming in for, for kind of chronic condition or long-term conditions and so we need a more extensive history about the patient. Often patients can give you this but having information like a full medical history is what I would expect to happen in an outpatient clinic. So, any background history for the patient any significant current and past medical problems. Any previous tests and procedures they have had, perhaps most recent blood tests that have had, the medications they are on. So essentially, I would require a quite comprehensive history in an outpatient clinic”*. Participant 1 also explained the difference between the need for information in different incidents and said: *“well clearly in an outpatient clinic the patient's conscious and able to answer questions. Outpatients, generally we access the previous medical record because it's likely that they have had previous visits, in particular, all of the laboratory results or outpatient discharge summaries and patient discharge summaries, radiology results, pharmaceutical dispensing, everything that would normally say on medical record”*.

The accuracy of categorizing the information required in out-patient clinics seems to be a difficult mission, due to the interrelations among medical conditions and their corresponding causes. Medical practitioners in out-patient clinics may require different information depending on patients' medical conditions and the purpose of each clinical visit. Therefore, it is more accurate to expose patients' information to practitioners in out-patient clinics upon requesting it.

d) Information that is required for research purposes (R)

Question 4 of the interview asked about information that could be removed from patients' records without affecting the usability of these records for research purposes. Participants answered this question as to what they thought could be removed without affecting the accuracy of the information for providing healthcare to patients. Most participants stated that the more information removed from any record the vaguer and less accurate the information would be. Some participants found it difficult to answer this question as they believed that it would depend on the goal of the research for what the information is being used, some information might be useful for certain research while other might not. However, there were few terms repeated in almost all answers that could be removed without affecting the accuracy of the information for research purposes, they were mainly the explicit identifiers of patients which include name, physical address, contact details, NHI number, and exact date of birth.

Despite the possibility of removing the exact date of birth from patients' records for anonymization, participants stressed the importance of having the "age" or "age groups" in each record for accurate research outcome. Participant 11 said *"A lot of hospitals require the address for zoning purposes but in terms of the emergency provision and research that may not be as relevant. And the date of birth is probably important because you do need that to tell the age of a patient and age is one of the most predictive factors in research"*. Participant 2 gave an example to how age is important in any patient record, she said *"for example, you might have negative T waves at a certain age and that's normal and then at another age, it needs to be positive. So, it would be extremely important to have the age in that instance"*.

The removal of the "date of birth" from patients' records affects the accuracy of research outcomes; however, this can be overcome by replacing the date of birth by a certain age group of patients. Participant 1 elaborated on the need for age to remain in patients' records and said *"I think the year of birth is important because we need to know the age of the patient so that we can get age-related population data. So perhaps you could remove the day and the month,*

but not the year of birth for children and infants you could remove the day but not the month because the difference between a six-month-old and 18-month-old is huge”.

4.3.2 Desired System Characteristics

For deriving the desired characteristics of healthcare information systems, participants were given a chance to criticize the current systems that they were using at work. They were asked to elaborate on how healthcare systems could be improved to get the best of what they could offer to improve the healthcare services provided to patients. Responses to all questions of the interview were used to identify how the healthcare system can best serve towards providing accurate healthcare to patients.

1. Information Storage

All participants explained the importance of storing all information about patients and their health. The storage of information was identified in participants’ responses to questions as one of the most important matters in the healthcare domain. Information about patients plays a vital role in the accuracy of healthcare provided to him/her. Participant 1 said, *“Well, I would say that all encounters with the health system should be recorded”*.

The storage of information in the current healthcare information systems was found in participants’ responses as a major challenge, not because information cannot be stored, but because of other issues related to storing it such as getting hold of the right information in the right time, and often different locations of storage places for the same patient. Participant 1 said, *“To improve the quality of health care provided to patients, the most obvious thing would be is access to the right information on the right patient at the right time”*. Participants explained that information about patients can be spread across various locations or holders. For example, a patient’s health-related information can be spread across various locations such as tertiary institutions like hospitals, primary like General Practitioners (GPs), and private healthcare institutions, this makes the process of obtaining all information related to the same patient difficult. Participant 13 in response to the last question in the interview about the desired characteristics of information systems in the healthcare domain said *“Well, I would say the answer to that is where you will need or you would like to have is one provider across New Zealand where you can have, and then you would get a direct link into every doctor into the hospitals, etc. One place and one provider of with the patients’ information instead of having so many as we have now at the present time we are”*.

The problem of not having all patients' information stored in one place creates a problem for medical practitioners to access important information about a patient being seen. Participant 15 gave an example of this problem and said *"The other thing that I feel is not well integrated as Auckland within Auckland is doing well but say if the patient was seen over in Hamilton then we don't kind of have access to anything. Or a patient was kind of seen in Wellington or they have kind of family live over there and spend half and half time here and there then that's actually really hard between cities. But there's no way if one of my patients got a blood test done over Hamilton I can't actually see it. I can't see a blood test or doctor's consultation that was in Wellington, yea Auckland area is getting a lot better integrated but intercity is not very much there"*.

Having a unified system for storing healthcare information seems to be a key solution for improving the healthcare domain. Despite the integration among various healthcare institutions in regard to connectivity and sharing information among each other, connectivity can often slow down the process of obtaining information from an institute to another. Participant 9 elaborated on this and said *"I think that this is difficult, but I think a single unified system or more unified platform would be useful that's easily accessible secure to everybody and that's that is fast. I mean currently, we have some issues with the speed of connective systems. I mean there are systems that talk to each other but there are quite slow that does impede performance, so I think speed and having a universal platform would probably be the most beneficial things"*.

2. Information Disclosure

The ability to get hold of information about patients did not seem to be the only challenge in the current health care information systems. Receiving the right information for the right patient at the right time is also a challenge. Several participants explained that getting information that is more what a medical practitioner requires in a certain incident can cause delays. The participant explained that even though everything should be recorded, not all details may always be needed. The disclosure of the right information contributes significantly to the accuracy and speed of healthcare provided to patients, participant 11 said in criticism to the current information system in the healthcare domain *"In addition to that I think the problem that we often have with the difference between hospital systems and our system is that the hospital system takes a lot of unnecessary information and publishes it on to their historic list and so often you have to scroll through a lot of old documents to get to the relevant section that you want"*.

Despite the difficulty in categorizing patients' records in terms of their importance, it was clear from participants that there are many encounters or information about incidents that do not influence future treatment for patients, for example, information about minor incidents such as skin tear and wounds may not always be needed, in fact, may never be needed in many cases, participant 1 elaborated on this and said *"Things that don't require to be stored, I'm thinking of when I review a patient's not an electronic medical record, their paper record would be things like the temperature chart and blood pressure during an operation for example. Those types of daily observations or urine output daily observations during an inpatient stay once a patient is discharged, they probably don't have much relevance. However, if the patient dies then they do have relevance. So, it just depends on the patient, the problem at the time"*. Another participant elaborated on this and explained that scrolling through un-needed information is annoying, especially when they are considered old information or non-updated.

The disclosure of the right information contributes to improve the healthcare provided to patients in terms of time and accuracy. Participants also explained that updating records of patients is an important matter; a suggestion made by one participant is to have a better way to categorize records and encounters so that practitioners do not have to scroll down through dates.

Sufficient disclosure of the information is derived as an important characteristic in a healthcare information system. Across responses from all participants, it was clear to notice the difference of information that each practitioner requires according to his/her role in the healthcare domain, however, disclosing the sufficient information can be a characteristic considered for this research. This can be associated with the dates of patients' medical encounters and impact on future treatment/presentation, latest update, and validity of information at the time of patient visit. Participant 11 elaborated on this issue and said *"a lot of the information we have in our system is different from the information that's available on the hospital system often Concerto or another platform. And when there's contradicting information there it ends up being that we have to take into account both of them. Now if there were a system where everybody could update in real-time with just one template at the end of a day that would reduce a lot of extra work"*.

3. Information Accessibility

Accessibility to information in the current healthcare information systems appeared to be a significant challenge. Some participants in their answers elaborated on the accessibility

challenge and explained that having different systems to deal cause difficulty in obtaining the information needed. The lack of standardization in how patients' information is presented and accessible creates a challenge when there is a need to access it. The majority of participants elaborated on the accessibility issue directly and indirectly in their answers, for example, participant 16 said *“For healthcare information systems, I think accessibility is pretty key, the hospital and general practitioners have access to each other's clinic letters, results, operation notes et cetera, that the template of the software looks similar in each setting so each time I'm a doctor may go to a different clinic. It's easy for them to use and familiarize themselves with”*.

Getting access to the right information in a time-efficient manner is a characteristic that helps improve the healthcare services provided to patients. Some participants explained that often information can be available to access for a particular patient but through different systems, and this causes delays. For example, participant 15 said *“something that happens kind of as soon as that's done any tests that were ordered from those places those test results come back to us as kind of the primary carers as well. So, it needs to kind of happen quickly as well as being easily accessible so either through the same system, the same system will be awesome but that's unlikely to happen or I guess easily through logins”*.

The ease of accessibility seemed to be an important characteristic in healthcare information systems to assure improvement in the healthcare services provided to patients in terms of accuracy and speed. This was concluded in participant 2 answer to the question regarding desired characteristics of healthcare information systems which were *“having a way that you could drill down and get more information as needed from a single platform, and yeah having it obviously be accurate and fast would be the things I think that would improve their health care quality for patients”*.

The desired system characteristics identified after the analysis of the collected data in this research are:

1. Just-Efficient information disclosure

Despite the importance of recording all information about patients and their health, information about patient health can be categorized according to their significance on future treatments/care to the patient and their needs in different encounters. Some information may not be needed in certain cases while other information is always needed. Disclosing the right information is a key characteristic of healthcare information systems to improve the healthcare services provided to patients.

2. Accessible Location of Information

Storing patients' information in once place is another key characteristic of the healthcare information system. Improving the ability to reach the right information about any patient starts with locating their storage.

3. Unified Platform

Accessing the information about patients through a single and unified platform is key to getting the right information about patients. The standardized platform user interface was a characteristic derived from the majority of respondents' answers. Figure (4.1) summarizes the findings of the case study research.

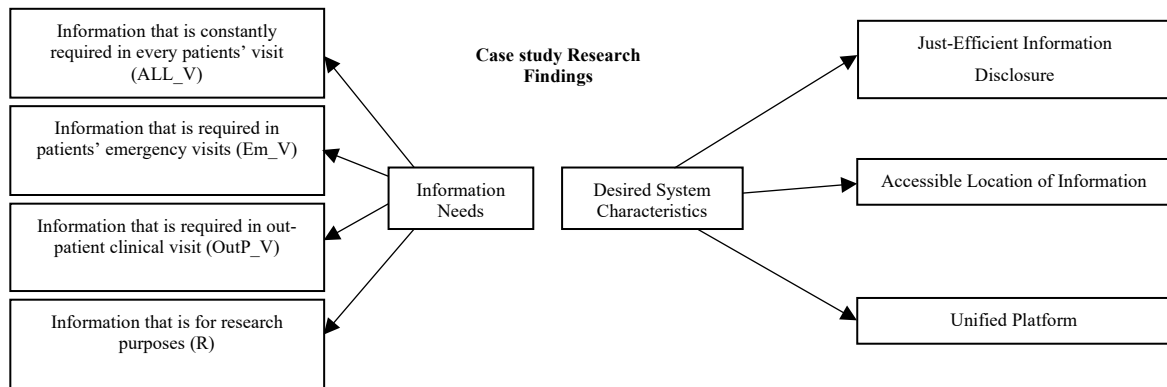


Figure (4.1) Case study findings

4.4 Summary

This chapter presented the entire process of collecting data, analyzing it, and reaching the findings. Section (4.1) presented how data was collected from research participants and organized for analyzing it, section (4.2) presented how the collected data was analyzed, section (4.3) presented the discussion of the analysis and how they can lead to certain findings, while section (4.4) presented the findings of the data analysis. The findings of the data analysis indicated desired characteristics that medical practitioners (users) desired to have in the information systems that they used. The findings also indicated the importance of information and how it is used for providing healthcare services to patients. The derived characteristics of healthcare information systems identified in this study were considered in the design of the target cloud-based architecture in this study.

Chapter 5: Proposed Cloud Architectural Design

Considering the complexity of the existing healthcare structure where patients' health information is distributed to multiple entities such as hospitals, healthcare centers, and cloud servers, a centralized architectural design of information systems for the healthcare domain would not be suitable, especially when interoperability remains a challenging obstacle among the vast majority of healthcare information systems. A non-centralized architectural design would be the most suitable option for the healthcare sector so that disparate entities can collaborate through sharing information related to patients and their health. This chapter presents a proposed cloud architectural design for storing patients' information and collaboratively using it concerning the privacy and confidentiality of it. The proposed design is different from the existing system designs because it allows for using the technology of cloud computing without risking the privacy of the information. As mentioned earlier, storing information in the hand of a third party (cloud provider) may lead to issues related to privacy. In the proposed design, it is possible to store information on the cloud without the ability of the cloud provider to learn its content, and only authorized users can access the information that is stored on the cloud. This is achieved by employing a searchable encryption mechanism to search through encrypted information while it is stored on the cloud, and categorizing patients' information according to the need of it in different contexts e.g. emergency instances, outpatient clinical visits, research ... etc. Moreover, the proposed design overcomes the challenge of interoperability by employing standard mechanisms in storing and accessing information in a privacy-preserving manner.

The proposed cloud architectural design is presented in two sections: the first section presents the proposed design for sharing healthcare information for medical treatment purposes in a privacy-preserving manner, and the second section presents the proposed design for using patient health-related information for research purposes without violating the privacy of individuals. The main issue in the adoption of cloud computing in the healthcare domain is keeping sensitive information in the hand of a third party. The owners of the data (patients' records) demand high levels of security on their data when they outsource it to a cloud. Although data is usually encrypted, whoever owns the data (e.g. patients, doctors, medical centers... etc) require having control over their data to perform operations such as updating records. In the normal process, data transferred to the cloud goes through traditional encryption methods for security reasons, however, the data holder needs to decrypt the data whenever an

operation is required on it. The data user provides the private key to the cloud provider to decrypt data to execute any required calculations. The decryption of the data at the cloud provider side causes privacy and confidentiality issues.

Moreover, a patient's record may include information that might not always be needed for all different instances of medical treatments, for example, a patient who has a certain sexual disease might not want a practitioner at an emergency practice to access and read information related to such disease when it is not needed in that particular treatment instance, therefore, accessing such unneeded information may also cause a breach of the patient's privacy.

In this thesis, a new cloud-based architecture is proposed for storing and sharing healthcare information in a privacy-preserving manner. There are two sources of information that informed the proposed design namely case study findings and literature review. The case study findings have fed the research with information related to: (1) how patients' health-related information is used for medical treatment purposes, and (2) the desired characteristics that healthcare information systems should have to best serve the domain in terms of storing and sharing information. The literature has fed the research with information that is related to the privacy requirements for healthcare information, and the potential privacy-related attacks that may be performed on patients' information.

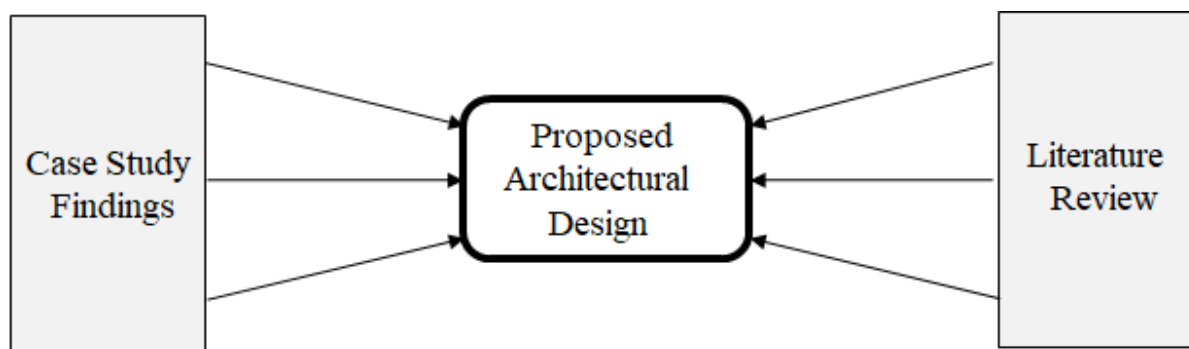


Figure (5.1) Information sources for the proposed architectural design

The information obtained from the case study findings and the literature has aided the process of designing the proposed architecture. The objective of the cloud architecture design was to: (1) satisfy the need for it from users' point of view (**case study findings**), (2) have the desired characteristics to best serve the healthcare domain (**case study findings**), (3) satisfy the privacy requirements related to storing and sharing healthcare information (**literature review**), and (4) meet the expectations of patients in regards to storing and sharing their health-related information (**literature review**). Table (5.1) presents the characteristics identified in the case

study findings and the review of the literature which were achieved in the proposed architectural design.

The proposed design of the cloud architecture includes two parts: the first part aims to store patient information on the cloud to collaboratively use it for medical treatment purposes, while the second part aims to use patient information for research purposes. The following section presents the fundamentals, components, and structure of the proposed architecture design for collaborative use of patient information for medical treatment purposes, followed by another section that presents the same for using patients' information for research purposes.

Target characteristic		Identified from
1	Just-enough information disclosure	Case study findings
	Disclosing only the right information according to the context in which information is required.	
2	Accessible location of information	
	Storing patients' information in one place for easy access whenever information is required	
3	Unified platform	
	Accessing information through a unified platform is a key characteristic toward improving healthcare services	
4	Adheres to the legal privacy-related frameworks	Literature review
	The architecture should adhere to privacy-related regulations and policies such as HIPPA and the information privacy act when using information.	
5	Patients control	
	Patients should have a means of control over who can access their information	
6	Cloud provider blindness	
	The cloud provider should not be able to read or access patients' information that is stored on the cloud	

Table (5.1) summary of identified system characteristics

5.1 Storing and Sharing Information

This section presents the fundamental aspects, components, and design of the proposed cloud architecture for collaboratively using patient health information for medical treatment purposes. The main objective of the design is to meet the six characteristics presented in table (5.1) of which three were derived from the findings of the case study and the rest from the literature.

There are two fundamental aspects of the proposed architecture design that enable it to store and share patient health-related information in a privacy-preserving manner. The first fundamental aspect is structuring patient information into categories. This aims to eliminate the exposure of information that is not needed during instances of medical treatments. Structuring patient information also contributes towards allowing patients to have means of control over who can access their information while it is stored on the cloud. The second fundamental aspect is the use of a searchable symmetric encryption scheme (SSE) which enables to search through encrypted information without decrypting it. The objective of the searchable encryption scheme is to store patient information on the cloud without the ability of the cloud provider to learn the content of the stored information. Further explanation of the objective of using the SSE scheme is provided later in this chapter.

5.1.1 Structuring Patients Information

A fundamental aspect of the proposed cloud architectural design is the accommodation of patients' health information under four main categories which were identified in the case study findings in Chapter 4. These categories are Information that is constantly required in every patients' visit (All_V), Information that is required in patients' emergency visits (Em_V), and Information that is required in out-patients' clinical visits (OutP_V), and information required for research purposes (R). The identified information categories were derived from the case study research. This section focuses on information categories that are used for medical treatment purposes, therefore, the (R) category is not included in this section.

The main goal of structuring patients' health information is twofold; firstly, to limit the exposure of information in instances when it is not needed. Disclosing just-enough information was identified as one of the important characteristics in the case study findings. For example, a medical practitioner in an emergency practice would need to access information about a patient that is required to perform accurate procedures for stabilizing the patient, while other

information such as information related to sexual health may or may not be required in the same incident. Secondly, limiting the exposure of information leads to better means of privacy protection that patients desire to have for their health information. Patients' information in the proposed system design is stored as a collection of files that are grouped into three different combinations referred to as documents.

Each document has an identifying tag and contains a number of files. Each file has the name of the patient, the name of the document that it belongs to, and a sub-tag used by the application system to identify it and locate it. The documents' identifying tags do not indicate anything

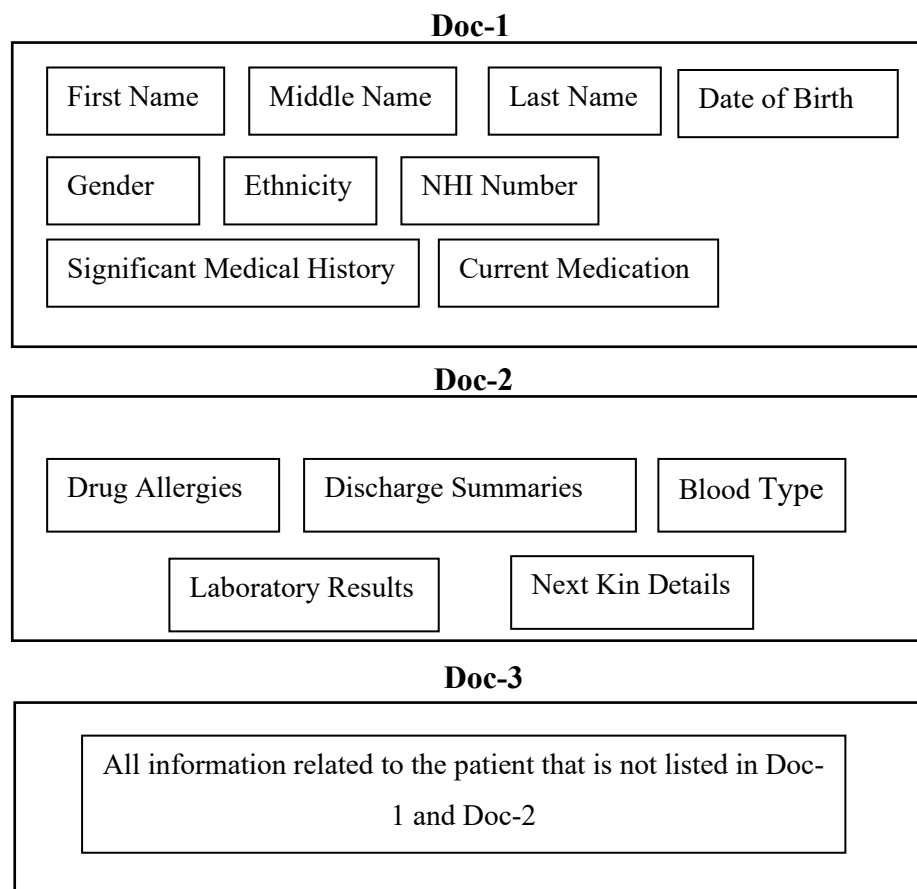


Figure (5.2) Individual documents that comprise patient information

about the content of their corresponding documents, they are used for technical purposes in the system and related to granting access to users. For example, and to simplify the description of the system design, the tags used for the documents are 1, 2, and 3. All patients registered in the proposed system have their information organized into doc-1, doc-2, and doc-3. Further explanation of how documents' tags are used is provided further in this section. Figure (5.2) presents the information contained in every document in light of the case study data analysis and findings. However, in the practical implementation of the proposed system, information

stored under each category is subject to change according to the medical treatment changing needs.

Individual documents and/or combinations of them comprise 3 information categories which are All_V, Em_V, and OutP_V which were identified in the findings of the case study research. Table (5.2) presents a description of the information contained in every information category.

Information Category	Information included
All_V	Patients' identifications, demographics details such as age and sex, current medications, and significant medical history and conditions.
Em_V	Along with (All_V): drug allergies, discharge summaries, Laboratory results, next kin details, and blood type.
OutP_V	All information upon requesting it. Accessing this information requires the patients' consent.

Table (5.2) Patients' information categories in the proposed system design

Every category comprises a number of files that exist in different documents. Each document contains different information related to the patient, therefore, accessing a category of information is a result of accessing one document or more. Figure (5.3 a) illustrates individual documents that contain files, while figure (5.3 b) illustrates a combination of all documents together making up the entire health information for a patient registered in the system. Files are tagged with system-generated numbers to be identified by the user application which is a component of the proposed system that is explained further in this chapter.

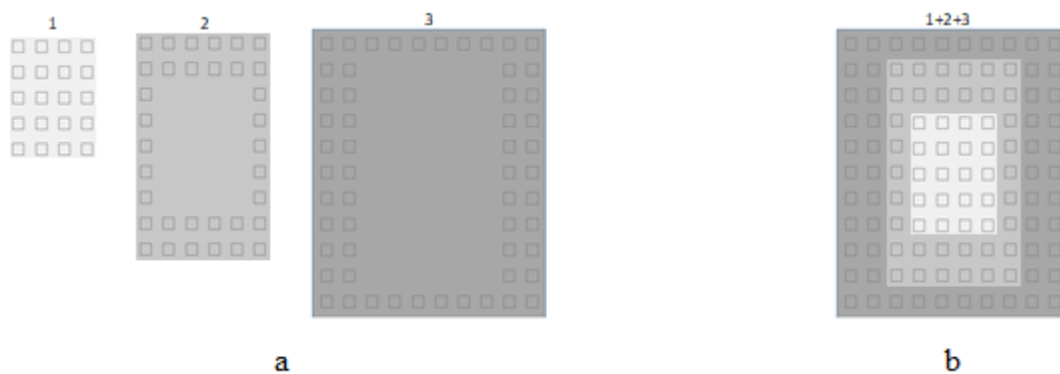


Figure (5.3) documents that contain files

The information categories comprise information contained in different documents as illustrated in figure (5.4). For example, when a user has the right to access information about a patient in an emergency setting (Em_V), doc-1 and doc-2 are released to the user, while a combination of document doc-1, doc-2 and doc-3 are released for users who have access to all information related to patients' health (OutP_V category). Tags of documents are used to identify them in the system. For every patient, the same tags are used for documents doc-1, doc-2, and doc-3. For example, when the Em_V category is requested, the system grants rights for accessing document doc-1 and doc-2.

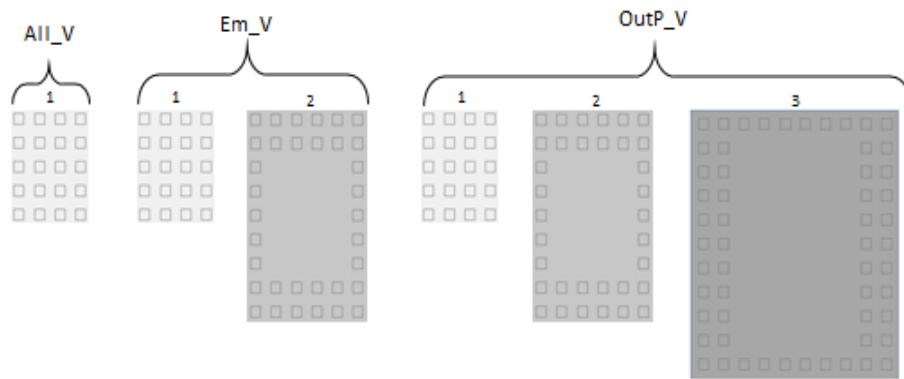


Figure (5.4) Information categories and their comprising documents

5.1.2 Searchable Symmetric Encryption (SSE)

Searchable symmetric encryption is a cornerstone of the proposed system architecture. The main objective of the proposed system architecture is to store patients' information on the cloud without the ability of the cloud provider to read it. Achieving this is considered easy but not practical without a mechanism that enables to search through encrypted information without decrypting it.

The proposed system employs a searchable symmetric encryption (SSE) approach (Curtmola, Garay, Kamara, & Ostrovsky, 2006). The SSE approach enables outsourcing data storage while preserving the ability to selectively search over it. There are three models for searching on encrypted data identified in the literature namely searching on public-key encrypted data (Boneh, Crescenzo, Ostrovsky, & Persiano, 2004), single-database private information retrieval (PIR) (Chang, 2004) and finally searching on private-key encrypted data (Curtmola, Garay, Kamara, & Ostrovsky, 2006) which is the approach employed in the proposed cloud architecture. For consistency purposes, the private key is denoted by secret key (S_k) throughout the thesis.

In the secret-key-encrypted data model, the data is encrypted by the user and is organized in an arbitrary way before encrypting it. The data is stored on a server in encrypted form and decrypting it can only happen using its S_k . In this model, the initial work for the user is large when data is large, while subsequent work such as accessing the data is small. The user work is large because data pre-processing requires performing several processes to facilitate searchability on it while it is encrypted. Structuring data as part of the pre-processing allows for efficient access to relevant data. In this proposed system, Information is partitioned into portions denoted by documents as explained earlier. For every patient, there is a root secret key (S_{KR}) that is used to encrypt 3 secret keys (S_k). Secret keys are used to encrypt patients' documents (doc-1, doc-2, and doc-3). Each document is encrypted with its corresponding S_k . Indexes and trapdoors -explained further in this section- are generated to identify and decrypt documents respectively. An important property of the secret-key-encrypted approach is that anyone who can decrypt information for a document can also decrypt any file in that document. This means anyone who has access to a document can have access to all files within that document.

The main goal of employing the SSE approach is to store patients' health information on the cloud in a searchable manner and only authorized parties can access it. Moreover, the cloud provider can never learn anything about the information stored, it receives encrypted information to store and releases it without decrypting it.

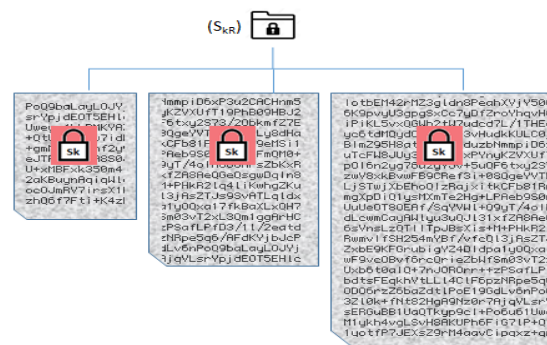


Figure (5.5) Encrypted structured patient's information

Structuring patients' information is key to the usefulness of the SSE approach. Figure (5.5) illustrates how patients' information is structured and how secret keys are distributed in the employed encryption approach. The decryption of each document under the secret root key requires the secret key for it which is released upon authenticated and authorized user requests. The cloud provider is not informed about the content of any document; therefore, the challenge

remains in identifying encrypted document/s without decrypting them. The searching capability of the SSE approach is achieved using a *secure index* mechanism (Goh, 2003). The secure index is a structure of data that stores document collections while supporting efficient keyword search, for example, given a keyword (w), the index returns a pointer to the documents that contain it. The secure index works by searching for a string exact match in encrypted documents. Every document contains a collection of encrypted strings, and a string is chosen to be the searching keyword for the document that contains it. The selected keyword is computed using the secret key by which the entire document is encrypted. The resulting ciphertext is then used to search for an exact match in documents. For example, a keyword in a document is “Basic-Information”. This keyword is computed using the secret key of the encrypted document and the resulting ciphertext is e.g. “JK^78Uo8361KL\$#VWL”. The combination of keyword and its corresponding ciphertext is then used to identify the document which contains the keyword “Basic-Information”. However, a keyword may appear in different documents, therefore, a number of keywords and their corresponding ciphertexts are put together in an encrypted index and corresponding trapdoor to assure the accuracy of document identification. Alternatively, a document’s unique name can be used to achieve the same

Secret Root Key (S_{RK})		Secret Root Key (S_{RK})	
Doc 1 - Encrypted Index		Doc 1 – Trapdoor	
Basic-Information	JK^78Uo8361KL\$#VWL	S_k	JK^78Uo8361KL\$#VWL
Significant	RM4%+B923SQLE@		RM4%+B923SQLE@
Current-Medication	XC%!OH2Nk_T[L7756LTFY\$#X		XC%!OH2Nk_T[L7756LTFY\$#X
Document-1	Y\$R<LB&&x2\$++J5		Y\$R<LB&&x2\$++J5

Table (5.3) Example of an encrypted index and Trapdoor for encrypted document outcome accurately such as doc 1. Table (5.3) demonstrates an example of an encrypted index generated for a document listed under a secret root key and its corresponding trapdoor.

To achieve the properties of the SSE approach, Curtmola et al (2006) proposed the below five algorithms which are the Key Generation algorithm (**KeyGen**), Key derivation algorithm (**KeyDer**), Index Generation algorithm (**IndexGen**), Trapdoor Generation algorithm (**Trap**), and a Search algorithm (**Search**). Below is the description of these algorithms:

KeyGen Algorithm

The KeyGen algorithm is a probabilistic algorithm that sets up the searchable encryption scheme. It is responsible for generating a secret root key for patient's documents as a collection. It takes a security parameter k and generates a secret root key (S_{KR}) for the patient S_{KR} . This key is used for wrapping and unwrapping the secret keys of all documents that belong to the patient.

$$KeyGen(1^k) \rightarrow S_{KR}$$

KeyDer Algorithm

The KeyDer algorithm is employed for generating a secret key (S_k) for each document listed under the secret root key (S_{KR}). It takes the document name and secret root key (S_{KR}) as input and generates a secret key (S_k) for the document. This secret key will be used to encrypt and decrypt the information contained in its corresponding document.

$$KeyDer(sk_{(i_1 \dots i_{n-1})}, (i_1 \dots i_n)) \rightarrow sk_{(i_1 \dots i_n)}$$

$$Fsk_{(i_1 \dots i_{n-1})}(i_n)$$

IndexGen Algorithm

The IndexGen algorithm is responsible for generating an encrypted index (I) for every document. It takes a number of keywords in a document such as the name of the document or its title and encrypts them using the document secret key (S_k). The output of the IndexGen algorithm is an encrypted searchable index I for every document to be used for searching it.

$$IndexGen(sk_{(i_1 \dots i_n)}, (i_1 \dots i_n), word_{w_1 \dots w_n}) \rightarrow sk_{(i_1 \dots i_n)}, I$$

$$Enc(sk_{(i_1 \dots i_n)}, I) \rightarrow C_1$$

Trap Algorithm

The Trap algorithm is responsible for generating trapdoors for documents. It takes the secret key of a document and keywords' ciphertexts as input and outputs a corresponding trapdoor (T) which is used for decrypting the document.

$$Trap(sk_{(i_1 \dots i_n)}, (i_1 \dots i_n), word_{(w_1 \dots w_n)}) \rightarrow T$$

Search Algorithm

The Search algorithm uses the decrypted index and the trapdoor for one document to find it. It takes the decrypted index and the trapdoor as inputs and identifies the encrypted document as an output.

$$\text{Search}(C_1, T_1) \rightarrow \text{Encrypted ciphertexts}$$

Similarly, in the proposed system, the process of preparing patients information for storage involves five steps:

1. Generating a secret root key (S_{KR}) for the patient.
2. Generating a secret key (S_K) for every document of patient information and choosing a keyword of each document.
3. Keywords are encrypted using their corresponding secret keys and the resulting ciphertexts are listed to form an encrypted index
4. Trapdoors are then created which involves combining the secret keys with the ciphertexts. Trapdoors will be used to identify and decrypt documents
5. Documents are encrypted using their corresponding secret keys

By following the above five steps, it becomes feasible to search for patients' documents while they are encrypted without having to perform decryption operations on them. Further explanation of how information is obtained from the cloud and decrypted is provided in the following chapter.

Note: The SSE scheme had many advancements to prevent from various security attacks since it was developed, however, for elaboration purposes, it was used in its initial state of the art. The real implementation of the system must consider employing an advanced SSE scheme.

5.1.3 Architectural Design and Components

The proposed architecture comprises five architectural components that are required for storing healthcare information on the cloud and collaboratively use it in a privacy-preserving manner. This section presents the design of the proposed architecture for sharing healthcare information for medical treatment purposes and provides details about its comprising components. Figure (5.6) illustrates the architectural design and the relationship of its comprising components.

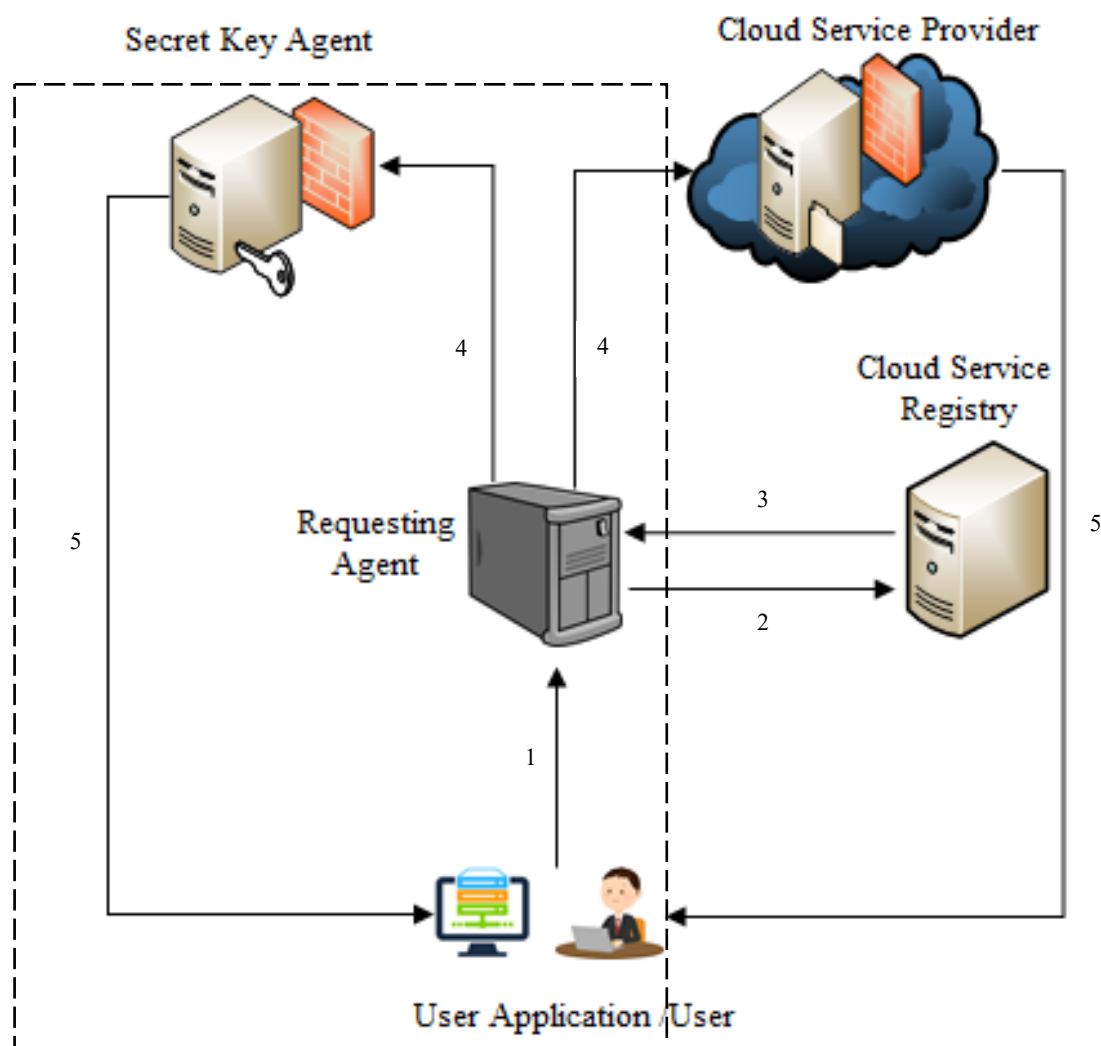


Figure (5.6) Proposed Architectural Design

As seen in figure (5.6), five components comprise the proposed architecture namely, Requesting Agent, User Application, Cloud Service Registry, Secret Key Server, and Cloud Service Provider. Each component is responsible to accomplish certain tasks as a contribution to achieving the main objectives of the proposed architecture. The components circled by the dash lines in figure (5.6) are the main contribution of this research.

Requesting Agent

The Requesting Agent (RA) is a server that is responsible for receiving requests from users and forwarding them to both the Cloud Service Provider and the Secret Key Agent after authenticating users. It is the point of contact through which users send requests to store or access information stored on the cloud. Users are authenticated and their access rights are identified before requests are forwarded by the RA. In other words, it plays the role of the gate person who does not allow unauthorized users to access the system. The RA has a limited communication channel with the users, a one-way communication channel with both the Cloud Service Provider (CSP) and the Secret Key Agent (SKA), and a two-way communication channel with the Cloud Service Registry (CSR) for users' authorization. The RA receives requests from users and only responds with information that is limited to confirmation of authentication. The one-way communication in the real implementation can happen by limiting the ability of the RA to respond to requests, responses must always be limited to acknowledgments or receiving requests from users through their user applications.

The RA stores the required information for identifying users and patients who are registered in the system. Information stored on the RA is important for facilitating secure access to patients' information that is stored on the cloud. Every patient has a unique code referred to as system ID which is generated by the RA and used for searching purposes.

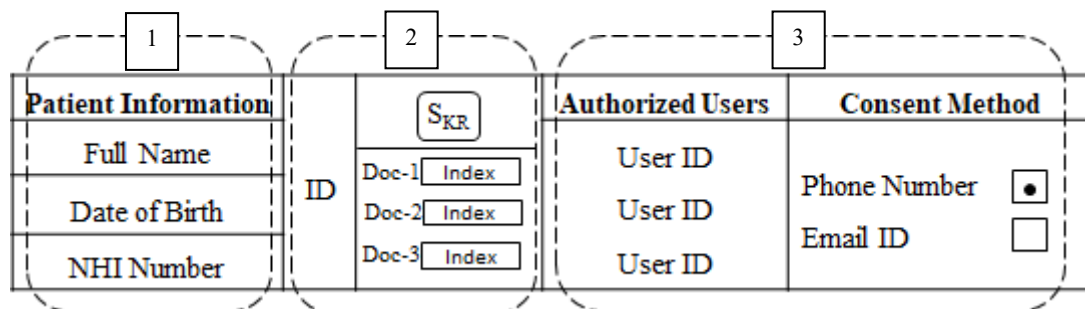


Figure (5.7) Information stored on the Requesting Agent for every patient

When a user requests to access patient information, the patient's system ID is used to identify the patient's information that is stored on the cloud. The main role of the RA in the proposed design includes receiving requests from users, authenticating users, and forwarding user requests to both the CSP and the SKA. The information stored on the RA for every patient is organized into 3 sections and every section contains information that is important to facilitate access to patients' information in a secure and privacy-preserving manner. Figure (5.7)

illustrates the information is stored in the RA. The dotted boxes represent the information sections numbered from 1 to 3.

Section 1 includes information that is required to identify patients on the system. Users request to access patients' information by using patients' basic identification information such as name, date of birth, and NHI number. Section 2 includes the patient's system ID, S_{KR} , and indexes. The system ID is required to identify patients' information that is stored on both; the cloud and the Secret Key Agent (SKA). The S_{KR} is required to decrypt the trapdoors that are stored on the SKA, and the indexes are needed to identify the documents stored on the CSP. Section 3 includes information that is required by the Cloud Service Registry (CSR) for authorizing users to access patients' information. It includes a list of users who have permanent consent to access the patient's information, as well as the patient's contact details for requesting and obtaining patient consent. There are two types of consents that patients grant to users for accessing their information: permanent consents which are granted by patients to their local GPs or pre-determined medical institutions/practitioners, and temporary consents which patients grant to medical practitioners/institutions for casual incidents or clinical visits. Patients optionally grant permanent consent to users to access their information. Temporary consent is granted to a user who does not have permanent consent and requires access to a patient's information. There are two methods of requesting and obtaining temporary consents in the proposed system: mobile phone in a form of text message, or via email confirmation. More information about patient consent is provided further in the following chapter.

Standard User Application

The proposed system architecture requires having a standard application that is installed and run locally on users' machines. Accessing patients' information stored on the cloud can only happen through a standard user application (UA). Having a unified platform to access patients' information was identified in the case study findings as a desired characteristic of healthcare information systems, therefore, the proposed architecture design employs a standard UA through which users can access information that is stored on the cloud.

The UA plays a key role in the proposed system architecture; it facilitates means of standardization to the process of storing, accessing, categorizing, and structuring information. There are three main functions that UA is responsible for which are storing, accessing, and updating patient information on the cloud. These functions are performed using buttons that

are available on the UA interface, these buttons are ENROL, REQUEST, UPDATE, and RESEARCH. Further explanation about these functions is presented in the following chapter.

There is a number of characteristics that UA has which enable it to store, access, and update patients' information on the system. The following are the main characteristics of the UA employed in the proposed system design:

1. Standard presentation and categorization of information

Categorizing information is part of the UA's functionalities. The application organizes patient information files into three documents (doc-1, doc-2, and doc-3) before it is stored on the cloud. The UA has a standard user interface for all users. Information is accessible when it appears in predetermined fields on the user interface. Information is presented in their associated fields only when it is decrypted. Information fields remain blank when their corresponding files are not decrypted. For example, a field on the application interface is predetermined for information related to patient mental health, this field remains blank when the logged-in user is not authorized to access the document in which mental health file exists.

2. Information pre-processing, encrypting and decrypting

The properties of the searchable symmetric encryption (SSE) approach employed in the proposed architectural design are achieved by operations performed by the UA. The pre-processing operations together with encryption/decryption operations are all performed by the UA. The UA is responsible to pre-process the information by organizing them and encrypting them following the SSE approach before it is sent for storage. It is also responsible for requesting to access information and decrypt it when it is received.

3. Characteristics related to accessing patients' information for research purposes

The UA has important characteristics that are related to accessing patients' information for research purposes in a privacy-preserving manner. Entering kiosk mode, disabling certain functionalities such as copy-paste functionality, allowing/prohibiting communication channels ... etc are all important characteristics of the UA. These characteristics aim to ensure the privacy of patients' information when used for research purposes. Further details about the characteristics related to using patients' information for research purposes are provided further in this chapter.

Cloud Service Registry

The proposed cloud architecture in this research employs the concept of the user identity management protocol for the cloud computing paradigm (U-IDM) proposed in (Eludiora, et al., 2011). U-IDM was initially proposed for cloud computing customers and cloud service providers. The main objectives of U-IDM were to achieve a set of global security objectives in cloud computing environments which include user authentication, authorization, and accounting. It aimed to protect customers and cloud provider's infrastructures by preventing unauthorized users to gain access to services or facilities delivered by cloud providers.

The main component of the U-IDM paradigm is the Cloud Service Registry (CSR). The CSR plays a vital role in the proposed architecture. It is located on the cloud side rather than on the client-side. CSR provisions access information according to users' privileges in a form of service level agreements (SLAs). Services in the context of the proposed architecture include the provision of access to patients' information that is stored on the cloud. Three information categories require access rights from the CSR which are All_V, Em_V, and OutP_V. As discussed earlier, each information category contains one or more documents. The CSR grants access to information categories by providing access to the documents that form these categories. Repeating the example of the Em_V category, it is a combination of doc-1 and doc-2 (Refer to figure (5.4) which represents the structure of information categories followed throughout this thesis). Therefore, granting access to the Em_V category requires the CSR to include the name of documents or their identifying tags with the user authentication confirmation. The CSR stores the names of categories and their comprising documents' tags. A list of registered users is stored on the CSR. Each user has a record of information related to

Org: AU43k1				
29930894	Nurse	Julia Robin	doc-1 doc-2	Em_V
29930804	Doctor	Steve Jordan	doc-1 doc-2 doc-3	OutP_V
29930814	Doctor	Ray Gather	doc-1 doc-2 doc-3	R
29930887	Assistant	Bob Parker	doc-1	All_V
29930842	Nurse	Andrew Keene	doc-1 doc-2	Em_V
29930832	Doctor	Robert Jameson	doc-1 doc-2 doc-3	Out_V

Table (5.4) Example of users list in stored on the CSR

their roles in the medical sector and the types of information that they can access. Users are listed under the name of their organizations. Searching for a user requires knowing the organization he/she belongs to. Table (5.4) illustrates an example of users' lists who are affiliated to an organization.

Nevertheless, an important task of the CSR in the proposed architecture is to obtain patients' consent for accessing their information. The CSR does not authorize users to access patients' information without having patients' consent. As mentioned earlier, when a user requests to access patient information, the RA authenticates the user and forwards the request to the CSR for authorization. Part of the information included in the RA's forwarded request includes a list of permanently authorized users to access the patient's information. This list enables the CSR to find out whether the user is granted permanent consent to access the patient's information or not. If the user is not included in the list, the CSR promptly sends a request for temporary consent to the patient, and the patient can promptly grant consent or reject.

Secret Key Agent

The Secret Key Agent (SKA) resides in a server that stores the required information for decrypting information stored on the cloud. As explained earlier, for every patient, there are 3 secret keys (S_k) listed under a secret root key (S_{KR}) which are used to decrypt 3 documents. All secret keys are stored together with trapdoors for all documents related to one patient (under one S_{KR}). The main functionality of the SKA is to receive requests from the RA and send the required trapdoors directly to the user. SKA has a one-way communication channel with the RA which is to receive requests, and a one-way communication channel with users to send secret keys, encrypted indexes, and trapdoors.

Cloud Service Provider

The cloud service provider (CSP) holds information related to patients' health. The main goal of the proposed architecture is to store all patients' information in one place which is the cloud. The CSP serves by storing and releasing encrypted information related to patients upon users' requests. The CSP has a one-way communication channel with the RA and a one-way communication channel with users. It receives requests from authenticated and authorized users through the RA and releases the required information in its encrypted form to users. Information stored on the cloud is contained in encrypted documents. The CSP cannot learn anything about the content of the documents stored. The cloud receives encrypted documents to store and release them to users without performing any decryption process on the documents.

The CSP employs a string match algorithm that aims to identify documents. Every patient has 3 identical encrypted documents that are labeled by the patient's system ID. Further explanation about the objective of the string match algorithm is provided further in the following chapter.

5.2 Information for research purposes

Releasing information for research purposes may lead to privacy breaches for patients in various cases. Individual patients may not wish to be individually identified when information about their health is disclosed and used for research. As found in the literature, the deletion of patients' explicit identifiers from their information such as name, address, and contact numbers is not enough for assuring that individual patients cannot be re-identified. There is a number of mechanisms identified in the literature by which attackers can identify individual patients in anonymized datasets and compromise their privacy. Various anonymization techniques identified in the literature aim to prevent the privacy of individual patients in aggregated health-related data from privacy attacks. The main idea is to transform datasets in a way that individual patients cannot be identified using the information in hand. However, the efficiency of these mechanisms is always at the cost of the integrity and usability of the anonymized datasets for research purposes.

The proposed architectural design aims to facilitate releasing patients' health-related information for research purposes in a privacy-preserving manner. It employs a research objectives-aware anonymization approach by which patients' information stored on the cloud is anonymized according to specified research objectives and required tasks. The following section presents a description of potential privacy breaches that can happen on patients' information when this information is aggregated and released for research purposes.

5.2.1 Privacy preservation strategies

As mentioned earlier, the proposed architectural design aims to facilitate using patients' health-related information for research purposes in a privacy-preserving manner. The privacy preservation is achieved in the proposed design by incorporating many settings of the UA to have privacy-protective characteristics, as well as performing anonymization processes on patients' information before releasing it, researchers. The main objective of the privacy approach employed is to prevent the attacks illustrated in figure (13) namely, linkage attacks and probabilistic attacks.

Prevention from linkage attacks

Linkage attacks as defined earlier aim to re-identify individual patients in anonymized datasets by combining the anonymized data with other available datasets. The performance of a linkage attack requires having a combination of datasets together to undergo linkage processes performed by purposeful algorithms such as algorithms presented in (Al-Mamun, Aseltine, & Rajasekaran, 2016) and (Ferguson, Hannigan, & Stack, 2018). Therefore, in the proposed system design, to perform a linkage attack on patients' anonymized dataset, the adversary needs to either (1) upload an external dataset onto the data analytics platform on the UA, or (2) download released datasets into different systems under its control.

The strategy adopted in the proposed system protects patients' information from linkage attacks by not allowing for combining patients' datasets with any other datasets. The main goal is to prevent researchers from performing any linkage attack on any released dataset. For this, several protective characteristics of UA are required which are listed below.

1. Application kiosk mode

Throughout the time of user access to research data, the user is not allowed to run any other application/program on the machine. It must be ensured that the user is locked into the user interface that only allows for requesting data and presenting it in a platform to locally run analytical queries for research purposes. This mitigates the ability to perform data linkage processes on the accessible dataset.

2. Communication channels

The UA must be configured for one communication channel with the RPS throughout the time of accessing patients' information. The goal is to block any other communication channel through which users can transfer datasets or part of datasets elsewhere or receive external information onto their machines.

3. Copy-paste functionality

The user interface must not support copy-paste functionality on datasets. There are various methods for copying data into external locations and have more means of control over it. Therefore, the UA should be configured to not support such functions. Moreover, the use of virtual input methods such as mouse and keyboard should not be supported by the user application when interacting with the RPS.

4. Activity Log

Users' actions on datasets and UA interface should be recorded. This enforces another line of security and privacy to patients' datasets. Users are held accountable for predetermined actions that are unpermitted.

Such configurations to the UA mitigate the ability of researchers to misuse patients' information by performing linkage processes on them. Users with the above system configurations and characteristics will not be able to combine the datasets released with other datasets. This mitigates the possibility of a linkage attack on patients' datasets.

Prevention from the probabilistic threat

A probabilistic attack occurs when an adversary can directly identify an individual patient by making successful speculations. This can happen by inferring potential fits between individuals and other randomized information to directly identify them. Direct identification can cause three privacy breaches namely attribute disclosure, identity disclosure, and membership disclosure. Insider curiosity and accidental disclosures are examples of direct disclosure without the researchers deliberately performing attacks (Appari & Johnson, 1997).

The proposed system employs a utility-aware de-identification approach to mitigate the ability of the researcher to violate the privacy of individual patients when using patients' datasets for research purposes, with preservation to the utility of the data in terms of its utility for research. The main objective of the approach is to ensure that an adequate degree of uncertainty is introduced when identifying data subjects. The computational operations for de-identifying healthcare data are performed by generalization hierarchies (El-Emam, et al., 2009)(Xia, Heatherly, Ding, Li, & Malin, 2015). The main goal of the generalization approach as explained earlier is to replace specific values with more general ones, and this results in many tuples of the dataset that have a similar set of quasi identifier values.

K-anonymity is a popular anonymization technique which was proposed in (Samarati & Sweeney, 1998). *K*-anonymity is a key concept that was initially introduced to address the re-identification risk through linkage attacks of anonymized datasets. A dataset is considered *k*-anonymous if every combination of quasi-identifiers occurs in at least *k* number of rows in the dataset. The technique involves transforming personal health information in patients' datasets to make it difficult for adversaries to identify individual patients in the dataset. *K*-anonymity is achieved through two techniques namely suppression and generalization (Wong, Li, Fu, &

Wang, 2006). Suppression involves the deletion or not releasing a value at all. Suppressed values are represented in asterisk (*). Generalization refers to the process of replacing a value with a less specific but semantically consistent value. Generalization is a process of modifying a specific value to a general value according to the predefined hierarchy(Kiran & Kavya, 2012). For example, zip code can 557841 can be generalized to 55784*, and zip code 55784* can again be generalized to 5578** and so on according to the predefined hierarchy. Similarly, the age attribute can also be generalized by replacing it with interval, for example, the age of 23 falls in the age range of (20-30). Figure (5.8) illustrates an example of hierarchical generalization for the attributes age and zip code in a dataset.

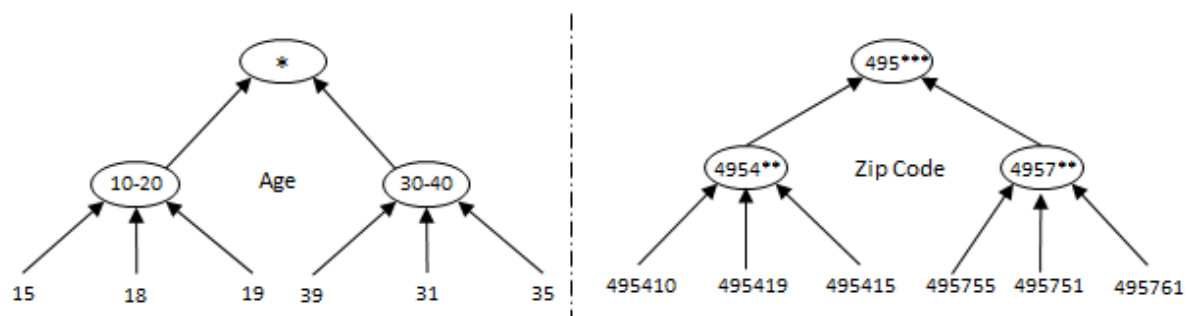


Figure (5.8) Hierarchical Generalization

A k -anonymized dataset has the property that each record is similar to at least k other records in the dataset, e.g. if $k = 2$ and the quasi-identifiers are age range, gender, and zip code, then the dataset has at least 2 records for each value combination of the age range, gender, and zip code. Records that do not meet the k requirement are suppressed entirely from the dataset. Table (5.6) presents a sample example of patients' 2-anonymity anonymized dataset.

Age Range	Gender	ZipCode	Condition
(30-40)	Female	522***	HIV
(30-40)	Female	522***	Breast Cancer
(10-20)	Male	522***	Heart Disease
(10-20)	Male	522***	Heart Disease
(50-60)	Female	522***	Diabetes
(50-60)	Female	522***	Colon Cancer

Table (5.5) Sample 2-anonymized dataset

As illustrated in Table (5.5), every tuple in the dataset that contains a similar combination of patient age, gender, and zip code appears 2 times in the dataset. This means any attempt to

identify an individual in the dataset will have at most $\frac{1}{k}$ accuracy. In k -anonymity, the higher the value of k leads to less identification accuracy. Such an approach is useful for preventing identity disclosure attacks.

However, despite the prevention from identity disclosure with k -anonymity, the privacy of patients' information is still prone to attribute disclosure attacks. K -anonymity does not provide sufficient privacy measures to patients' information. To explain the attribute disclosure attack, we assume that an adversary has the dataset presented in table (5.7) and another dataset obtained from an external source such as the government voting system. The combination of both tables (medical dataset and voters' dataset) may enable the adversary to associate sensitive attribute/s to a group of individuals without identifying them individually, this attack is called Homogeneity Attack. Table (5.7) presents how Attribute disclosure can occur using external datasets obtained from other sources.

External information				Anonymised medical dataset			
Name	Age	Gender	zip code	Age Range	Gender	ZipCode	Condition
Amanda	32	Female	522475	(30-40)	Female	522***	HIV
Sophie	39	Female	522652	(30-40)	Female	522***	Breast Cancer
Brandon	18	Male	522315	(10-20)	Male	522***	Heart Disease
James	15	Male	522989	(10-20)	Male	522***	Heart Disease
Ella	55	Female	522444	(50-60)	Female	522***	Heart Disease
Jennifer	51	Female	522444	(50-60)	Female	522***	Colon Cancer

Table (5.6) Attribute disclosure using external information

As seen in Table (5.6), the adversary may not be able to accurately identify Brandon or James but can find out that both of them have heart disease conditions. Such finding is a breach of the privacy of both Brandon and James. Therefore, k -anonymity suffers from the inability to protect the privacy of sensitive attributes due to the lack of diversity in the sensitive attributes. Another potential attack on k -anonymized datasets is the Background knowledge attack. Background knowledge attack happens when an adversary knows background knowledge about an individual and uses it to eliminate possible values for the sensitive attribute of a patient. Suppose an adversary knows that Alice is 35 years old, female, writer and has been to the hospital which published the table. The adversary can see that all the female writers of age 35 suffer from a common disease which is HIV. The attacker can then conclude that Alice suffers from HIV disease. This attack is also known as a positive disclosure attack (Manta, 2013).

In response to the issue of lack of diversity of sensitive attributes that k -anonymity has, the authors in (Machanavajjhala, Gehrke, Kifer, & Venkatasubramanian, 2006) proposed a new privacy definition called ℓ -Diversity which is adopted in the proposed system design. ℓ -Diversity overcomes the limitation of the k -anonymity model by providing privacy to the released dataset. This privacy protection is achieved by satisfying the requirement that the values of sensitive attributes are well-represented in each group. The technical concept of ℓ -diversity is a modification of k -anonymity by incorporating the k -anonymity principle (Li, Li, & Venkatasubramanian, 2007). The authors in (Machanavajjhala, Gehrke, Kifer, & Venkatasubramanian, 2006) state that for creating an algorithm for ℓ -diversity requires taking any k -anonymity algorithm and make the following change: every time a T^* table is tested for k -anonymity, the ℓ -diversity is checked instead.

To achieve the property of ℓ -diversity on a released dataset, every block of the released dataset must agree to the function $r_i < c(r_1 + r_{1+1} + \dots + r_m)$ where c refers to a constant parameter and r_i denotes the repetition of the sensitive value appearance in the block.

The main principle of ℓ -diversity as explained in (Machanavajjhala, Gehrke, Kifer, & Venkatasubramanian, 2006) is that a q^* -block is ℓ -diverse if it contains at least ℓ “well represented” values for sensitive attribute S . A table is ℓ -diverse if every q^* -block is ℓ -diverse as shown in table (5.8).

Age Range	Gender	Zip Code	Medical Condition
30-40	Female	521 ***	HIV
30-40	Female	521 ***	Colon Cancer
30-40	Female	521 ***	Diabetes
30-40	Female	521 ***	Diabetes
40-50	Male	522 ***	Diabetes
40-50	Male	522 ***	Colon Cancer
40-50	Male	522 ***	HIV
40-50	Male	522 ***	Diabetes

Table (5.7) 3-Diverse released patients table

As presented in Table (5.7), each block of the dataset contains 3 well-presented sensitive attributes namely HIV, Colon Cancer, and Heart Disease. This makes it more difficult to disclose an attribute that is associated with an individual patient within each block. Disclosing an attribute for any individual patient requires having ℓ pieces of background information. This eliminates the possibility of attribute disclosure for each individual.

However, despite the protection from attribute disclosure in the example illustrated in Table (5.7), an adversary can still violate the privacy of individuals by speculating with high confidence a range of attributes that can be potentially associated with an individual. Let's assume Alice knows that Bob is 42 years old male who lives in zip code 522112 and hence knows that his record exists in Table (5.6). Alice wants to find out whether Bob is HIV positive or not. She cannot reach such a conclusion with high confidence; however, she can be confident that Bob suffers from either HIV or diabetes or colon cancer. Moreover, an adversary with background knowledge can increase the risk of identifying patients by direct disclosure attacks namely positive disclosure and negative disclosure attacks explained in (Loukides & Shao, 2011). Positive disclosure as described earlier is when Alice knows that Bob suffers from colon cancer, this would enable her to identify Bob's record and infer other information contained in his record. Negative disclosure is when an adversary knows that the target individual does not suffer from a certain disease. For example, when Alice knows that Bob does not have diabetes, then she can infer with 50% confidence that Bob in Table (5.6) suffers from either HIV or colon cancer. Another risk example can be when a particular medical condition is associated with one gender such as breast cancer. When gender is not revealed in a dataset and the medical condition is breast cancer, then Alice can narrow down her speculation with high confidence by excluding tuples of the dataset that can potentially be Bob's.

To overcome the issue of positive disclosure, the authors in (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2006) further introduced a privacy property to ℓ -diversity called Recursive (c, ℓ) -Diversity. The main idea of this property is to control the occurrence frequency of attribute values in every q^* -block in a released dataset. In medical datasets, some positive disclosures of values are considered acceptable, because its disclosure does not violate privacy such as the value "Healthy" of the "Medical Condition" attribute. Recursive (c, ℓ) -diversity model aims to mask sensitive values by predefining sufficient occurrence frequency distribution for values in every q^* -block. The explanation of (c, ℓ) -diversity is as the following:

Let Y denote the set of sensitive values for which positive disclosure does not violate the privacy of patients in a dataset. In a given q^* -block, let the most frequent sensitive value that is not in Y be the y^{th} most frequent sensitive value. Let r_i denote the frequency of the i^{th} most frequent sensitive value in the q^* -block. The q^* -block satisfies the property of recursive (c, ℓ) -diversity if one of the following holds:

$$y \geq \ell - 1 \text{ and } r_y < c \sum_{j=\ell}^m r_j$$

$$y \geq \ell - 1 \text{ and } r_y < c \sum_{j=\ell-1}^{y-1} r_j + c \sum_{j=y+1}^m r_j$$

To explain the idea of Recursive (c, ℓ) -Diversity, suppose there are two values for the attribute “HIV”, positive and negative. The disclosure of the value “negative” does not violate the privacy of the patient while the disclosure of the value “positive” does. Therefore, to avoid positive disclosure of the attribute HIV “Positive” in the anonymized dataset, the frequency of the “positive” attribute occurrence is controlled. Such control to the frequency of occurrence happens by predefining the value of the parameter c . In recursive (c, ℓ) -diversity, the number of patients who are HIV “positive” appears in q^* -block is less than c times the number of HIV “negative” patients or, in other words, the number of HIV “positive” patients in the q^* -block is at most $\frac{c}{c+1}$ patients. For example, if c is predefined to be 0.04, then at most 4% of the patients in the q^* -block are HIV positive, and if c is predefined to be 1, then the maximum number of patients who are HIV positive in the q^* -block is calculated as $\frac{1}{1+1}$ which equals a half.

Controlling the occurrence frequency of values of the same attribute eliminates the possibility of positive disclosure attacks on anonymized datasets and therefore eliminating the chances of possible direct disclosure attacks. The recursive (c, ℓ) -diversity is a cornerstone in the proposed interactive utility-aware anonymization approach for releasing patients’ information for research purposes.

The proposed system design employs an interactive utility-aware anonymization approach for releasing patients’ information for research purposes, with considerations to research objectives and tasks. Information released to researchers is the only information required by them in their user requests. Further description is provided further in this section. The proposed system enables releasing patients’ records that satisfy the need of the records in terms of research, as well as eliminates the possibility of direct attribute disclosure that researchers may achieve intentionally or unintentionally. To further explain how the recursive (c, ℓ) -diversity model serves to eliminate the possibility of direct disclosure in anonymized datasets, Table (5.7) is referred to.

As seen in Table (5.7), an adversary can infer information about an individual by having some background knowledge. The example of Alice shows that she could conclude with high confidence that Bob suffers from either HIV, diabetes, or colon cancer just by knowing that Bob is a male and lives in zip code 522112. Such conclusion is achieved directly by Alice because every q^* -block in the released dataset contains only attributes that violate the privacy of individuals when associated with them. For this, the recursive (c, ℓ) -diversity mechanism aims to eliminate the confidence of researchers in speculating any association between individuals and attributes.

To eliminate the confidence of adversaries in speculating associations between attributes and individuals, more tuples that contain non-useful-but-correct values are released with predefined occurrences in every q^* -block in the published dataset, e.g. the value “Healthy” for the attribute “Medical Condition”. Table (5.8) presents an example of anonymized q^* -blocks using the concept of (c, ℓ) -diversity model.

As seen in Table (5.8), every q^* -block contains 2 tuples that contain the value “Healthy” in the medical condition attribute. This means if Alice wants to find out what medical condition

Age Range	Gender	Zip Code	Medical Condition
30-40	Female	521 ***	HIV
30-40	Female	521 ***	Colon Cancer
30-40	Female	521 ***	Diabetes
30-40	Female	521 ***	Healthy
30-40	Female	521 ***	Diabetes
30-40	Female	521 ***	Healthy
40-50	Male	522 ***	Diabetes
40-50	Male	522 ***	Healthy
40-50	Male	522 ***	Colon Cancer
40-50	Male	522 ***	Healthy
40-50	Male	522 ***	HIV
40-50	Male	522 ***	Diabetes

Table (5.8) (c, ℓ) -diverse dataset sample

Bob has, her confidence in making any speculation will be low. The conclusion that Alice can reach is that Bob may suffer from either diabetes, colon cancer, HIV, or he might be healthy.

The confidence of any speculation that Alice makes is calculated by dividing the number of occurrences of a sensitive value on the total number of tuples in the q^* -block. Therefore, Alice would need more pieces of information about Bob to directly identify his record in the anonymized dataset. For example, in the table (5.8), the possibility that Bob suffers from diabetes is calculated by dividing the number of tuples that have the value “diabetes” in the medical condition attribute by the total number of tuples in the entire q^* -block, which is $\frac{2}{6} = 0.33$. Therefore, controlling the occurrence of sensitive values in every q^* -block contributes significantly towards reducing the possibility of direct attribute disclosure.

The interactive nature of the proposed system design allows for efficient anonymization mechanisms concerning the research objectives for which the data is used. The term “efficiency” refers to the elimination of an adversary’s ability to directly identify an individual patient or associated attribute by using background knowledge about the patient. The data in the proposed system design is anonymized and released according to queries received by researchers.

Researchers tend to use patients’ records for certain goals that are normally specified in their research objectives. For example, if research requires accessing patients’ records for analyzing certain types of cancer, researchers for this would require access to records of patients who suffer from cancer. They also may be disparate to access other information that is necessary for the completeness of the information required e.g. age, sex, previous medical history ... etc. However, information about patients is presented as values of attributes in patients’ records, therefore, the proposed system design requires researchers to specify the objectives of their researches and the attributes required in the released patients’ records. The sensitive attributes required by researchers are referred to as the target attributes in datasets. Such information when included in researchers’ queries for accessing patients’ records helps the data holder significantly in anonymizing the data sufficiently concerning the utility of it for the intended research.

However, since it is not possible to predetermine standard purposes of researches for which patients’ records will be needed, the proposed system performs many operations on patients’ records prior to releasing them according to the intended use of them for research purposes. These operations aim to maintain the anonymity of patients’ information and eliminate the chances of direct disclosure that researchers may achieve when using anonymized datasets. The following section presents an overview of the proposed system design in terms of its

components, as well as operations performed on patient's information before releasing it for research purposes.

5.2.2 Architectural Design and Components

After discussing the strategies and techniques incorporated for sharing healthcare information for research purposes in a privacy-preserving manner, this section aims to present the design of the proposed architecture in terms of its components and their functionalities. Figure (5.9) presents the proposed architectural design for storing and using patients' information for research purposes. The components circled by the dash lines in figure (5.9) are the main contribution of this research.

The proposed architecture design for storing and using patient information for research purposes comprises four components namely: Research Portal Server (RSP), User Application (UA), Requesting Agent (RA), and finally the Cloud Service Registry (CSP). Each component plays its role in assuring patients' information is used for research purposes in a privacy-preserving manner.

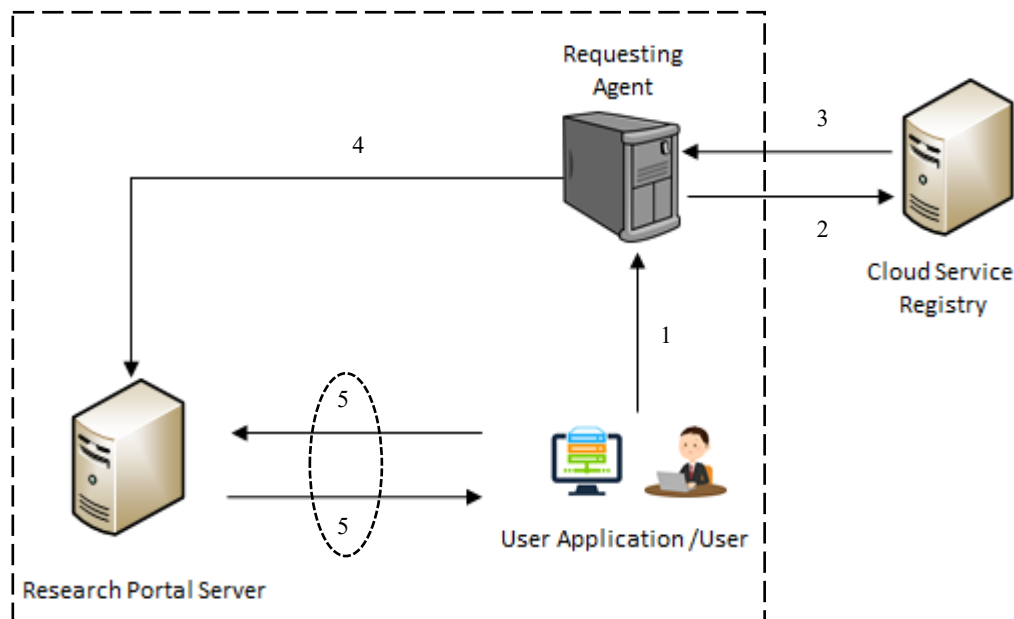


Figure (5.9) Architectural design for sharing healthcare information for research purposes

Research Portal Server

The Research Portal Server (RPS) is a key component of the proposed architecture that accommodates patient information for research purposes (R). It is a cloud-based server that

stores patient information in a standardized form and releases it for research purposes in an interactive manner.

The role of the RPS in the proposed architecture is vital; it is responsible for receiving queries from authenticated users and accordingly releasing information in the anonymized form concerning the intended use of the information. The RPS performs a number of operations on patient information prior to storing it which are explained further in this section. It also employs many anonymization algorithms that aim to anonymize patient information before releasing it to users.

User Application

The proposed system architecture requires having a standard user application (UA) that is installed and run locally on users' machines. Accessing patients' information stored on the RPS can only happen through the UA. The UA has several protective characteristics (discussed earlier) to prevent from possible linkage attacks that researchers may perform on released datasets. These characteristics are the ability to run in kiosk mode, blocking all communication channels except with the RPS, disabling copy-paste functionalities, and finally having an activity log to track users' actions during the use of patients' information.

Requesting Agent

The Requesting Agent (RA) is a server that is responsible for receiving requests from users (researchers) and forwarding them to the RSP. It is the point of contact through which users can connect to the RPS. Users are authenticated and their access rights are identified before requests are forwarded by the RA to the RPS. The RA has a one-way communication channel with the users, a one-way communication channel RPS, and a two-way communication channel with the CSR registry for users' authorization.

Cloud Service Registry

The Cloud Service Registry (CSR) plays a vital role in the proposed architecture. It provisions access to RPS according to users' privileges (SLAs). Users who are registered in the system as researchers are allowed to access the RPS.

5.2.3 Storing patients records on the Research Portal Server

Through the user application, when updating patient information or enrolling a subscriber in the system, there is an option that is available on the UA interface whether the patient allows

for using his/her information for research purposes or not. If consent is obtained, the option is selected and the patient's information is not only sent to the cloud for storage as explained in the previous section, but also to the RPS to be used for research purposes. However, information that arrives at the RPS undergoes many operations before storing them. These operations aim to organize patients' information in a standard form to facilitate releasing it for research purposes in a privacy-preserving manner. The reason for storing patients' information in the RPS is to achieve the following characteristics, also to further enable efficient anonymization features which are explained further in the following subsections.

1. Standardizing information presentation

As explained earlier, patient information is contained in three documents, each document contains many files, and each file contains certain information that is related to the patient's health.

All information contained in files is stored on the RPS as values of attributes that are standard for all the patients' information. Patients' information stored on the RPS is organized into rows in a "mother dataset" referred to as (R). All rows of the mother dataset have similar attributes that vary in values from a patient to another. For example, all rows have the attributes Medical Conditions, Medical Condition Severity, Blood Type, Age, Zip Code ... etc. When a patient's information arrives at the RPS, a new row in the mother dataset is created, and information contained in the files are distributed as values into their corresponding attributes.

2. Attributes suppression

In every patient's information, several attributes are needed for identification purposes, healthcare management, and other matters such as insurance. These attributes may include explicit identifiers and other information that is helpful for adversaries to identify individuals by having little background knowledge, such as occupation, an insurance company, employer, next kin, family doctor name ... etc. Such attributes do not affect the utility of the patient information for health-related research purposes when suppressed from the record. Therefore, attributes that hold information that is not related to the patients' health are entirely suppressed from patients' records before storing them. Suppressing attributes is part of the UA's functionalities. The UA sends only information that is useful for research purposes to the RSP and suppresses the rest.

5.2.4 Releasing datasets for research purposes

Upon receiving queries from researchers for having access to patients' information for research purposes, there are several operations that are performed by the system to ensure releasing a useful anonymized dataset that satisfies the need for it for each intended research.

Operation 1: Generating Dataset

Although researches vary in purpose and required attributes in patients' records, two main requirements should be considered when generating the dataset. These requirements are related to the size of the dataset (number of tuples) released, and the number of records that have the target attributes' values. These requirements aim to eliminate the possibility of direct disclosure without affecting the usability of the dataset. Researchers are required to provide information about the purpose of the research and the required attributes. It is part of the RPS's functionality to require information from researchers about the intended use of the information and the required attributes in the required datasets. Having a large number of tuples in the released dataset is required to assure that there are enough tuples in the dataset to effectively perform the anonymization processes, while the percentage of tuples that contain sensitive values aims to assure that there are enough tuples in the dataset that contain not-useful-but-correct values. The thresholds (number of tuples and percentage of tuples with sensitive target values) may not always be firm; they are configurable according to the objectives of the intended research, however, in this thesis, the following requirements are considered for instantiation purposes.

Requirement 1: No less than 3000 records are released in any dataset. This requirement aims to assure that there are enough tuples in the generated dataset to effectively apply the employed anonymization model. The main objective is to prevent potential direct disclosure threats explained earlier.

Requirement 2: The sensitive value of the target attributes should appear no more than 50% of the total tuples in the generated dataset. For example, in a dataset that contains 3000 tuples, if the value of the target attribute "Medical condition" is "lung cancer", there should be at least 1500 tuples that have other values of the same attribute. In other words, if a dataset is required for analyzing types of cancer, tuples that have the value 'cancer' or other cancer-related conditions for the medical condition attribute are required. Therefore, to satisfy the requirement that no more than 50% of tuples have the value "cancer" or other cancer-related condition, it is required to have at least 50% of the selected tuples that have the value of the medical condition

attribute ‘non-cancer’ or ‘non-cancer-related’ conditions such as “Healthy” or any other medical condition such as “HIV”.

Operation 2: Anonymization

After a dataset is generated, the RPS performs a number of processes that aim to anonymize the identity of individual patients and mask their sensitive attributes. The anonymization stage aims to assure that the patients’ information is anonymized prior to releasing it for research purposes. The anonymization operation in the proposed system involves two processes namely value generalization and value transformation.

1. Value Transformation

The process of transforming values aims to make sure that there are enough tuples in the dataset that have non-useful-but-correct values for the target attributes. Using the example of the cancer types analysis research, the values of attributes in tuples that have a different medical condition such as “HIV” or “Healthy” are transformed into a value that is not useful in context but correct in fact, such as “no cancer”. If the target values are a number of medical conditions, the non-useful-but-correct values can be “none” which is for patients who do not suffer from the specified medical conditions. Such transformation will result in a dataset that has at least 50% of tuples with values in the target attributes that do not violate the privacy of individuals when associated with them. The appearance of the tuples that have transformed values in the target attributes decreases the chance of successful speculations that a researcher may make to directly identify an individual or associate attribute to individual patients. Tuples with transformed values are excluded internally by the UA in any analytical operations performed during the research.

2. Value generalization

The main objective of the generalization process is to replace specific values of quasi-identifiers with more general ones. The goal is to achieve numbers of tuples in the dataset that have similar quasi-identifiers such as age, zip code, gender ... etc. This eliminates the chance of identifying individual patients/subscribers in released datasets. However, since the proposed system design is interactive, researchers are required to select attributes that are required in the requested datasets. The generalization process employed in the proposed system design is performed on the zip code value, date of birth (age) value, and gender value.

Zipcode

The zip code often has importance for the completeness of information in various research contexts, therefore, instead of suppressing it for anonymization purposes, it is generalized to different levels in the proposed system design as explained in figure (5.9). The need of zip code in patients' information is derived from the importance of the geographical areas in which patients live in, therefore, researchers are required to include in their queries the importance of the geographical area attribute (zip code) for their research, and accordingly, the zip code is generalized. The less importance of the zip code leads to a higher level of generalization. For example, the zip code may be anonymized by masking a number of characters to indicate a large geographical area such as an entire city or state.

Date of Birth

The date of birth attribute was identified in the data analysis phase of the research as an important factor in healthcare-related researches. Therefore, the deletion of such attributes from patients' information affects the utility of the information when used for healthcare-related research purposes. However, the need for the date of birth stems from the fact that age is an important factor in various matters related to health conditions and treatments. Medical practitioners when interviewed stated that replacing the date of birth with age interval sustains the utility of the information when used for research purposes. They indicated to certain age intervals that are useful to represent the age, these intervals are as the following:

[5 years range] for individuals who are 3 years old or above e.g. Age [3-8] years

[3 months range] for individuals who are below than 3 years old e.g. Age [12-15] months

Gender

Deriving from the literature and the collected data from medical practitioners during the data collection phase of this research, the gender is often an important factor in various health-related issues in terms of types of diseases or medical conditions that patients may have. Some diseases may only be associated with females while others to males, also it was confirmed a gender may be prone to a certain disease more than the other. Therefore, generalizing the gender to have (male and female) may often affect the utility of the records for research. However, since the proposed system design is interactive, researchers are required to indicate in their queries if the gender attribute is required, and the anonymization then can be on the cost of the generalization level of the age and the zip code in the released dataset. If gender is

required to be specific, a higher level of age and zip code generalization takes place. The more tuples of similar quasi-identifiers the less chance direct disclosure of individuals in the dataset.

Operation 3: Releasing (c, ℓ) -diverse dataset

The concept of recursive (c, ℓ) -diversity is employed in the system design. The rule is: every q^* -block in a released dataset must include tuples with values that do not violate the privacy of individuals when disclosed. The parameter c refers to the frequency of occurrence of the privacy-violating values of the target attributes in every q^* -blocks, while the ℓ represents the number of medical conditions such as unhealthy medical conditions that occur in the same q^* -block. Table (5.9) represents an example of q^* -blocks that satisfy $(2, 2)$ -diversity.

Age Range	Gender	Zip Code	Medical Condition
30-40	Female	521 ***	HIV -1
30-40	Female	521 ***	HIV -1
30-40	Female	521 ***	HIV -2
30-40	Female	521 ***	HIV (Negative)
30-40	Female	521 ***	HIV (Negative)
30-40	Female	521 ***	HIV (Negative)
40-50	Male	522 ***	HIV (Negative)
40-50	Male	522 ***	HIV -1
40-50	Male	522 ***	HIV (Negative)
40-50	Male	522 ***	HIV (Negative)
40-50	Male	522 ***	HIV -2
40-50	Male	522 ***	HIV -1

Table (5.9) $(2, 2)$ -diverse HIV patients block

As seen in table (5.9), there are a number of patients who are diagnosed with HIV positive. The percentage of tuples with HIV positive is 50% in each q^* -block, while the other 50% of tuples in every q^* -block is for patients who are HIV negative. Each q^* -block in the dataset contains tuples of both HIV positive and negative. Therefore, the chance of successful speculation to directly identify an individual or an attribute is low. For example, if Bob's record exists in the released dataset, there is 50% possibility that he is HIV negative, while the possibilities of him being HIV-1 and HIV-2 is $\frac{2}{6}$ and $\frac{1}{6}$ respectively. Therefore, when the number of tuples increases in every q^* -block in the dataset, the chance of successful direct disclosure decreases.

5.3 Summary

Employing cloud computing technology in the healthcare domain grants significant benefits in terms of information sharing. However, due to the number of challenges described earlier in this chapter, the adoption of cloud computing technology for healthcare information systems has always been limited. As discussed in this chapter, the proposed architectural design enables the healthcare domain to obtain the benefits of cloud computing technology by incorporating a number of strategies and techniques that overcome the major challenges related to the privacy of information. The proposed architecture design benefits the healthcare domain by facilitating collaborative use of information for both, providing healthcare services to patients and for research purposes. It enables to (1) store sensitive information on the cloud without the ability of the cloud provider to read it; (2) share patients' information for medical treatment purposes in a privacy-preserving manner; (3) grant patients means of control over who can access their health information (4), and finally use patient health-related information for research purposes without compromising the privacy of individual patients. The architecture also adheres to the main privacy requirements and legal frameworks outlined in Chapter 2. The adoption of the searchable encryption scheme and the separation of information (encrypted data and secret keys) enable the proposed architecture to prevent the privacy threats that could be performed due to the ability of the cloud provider to read it. The user identity management protocol (U-IDM) preserves the confidentiality of the information that is stored on the cloud and grants patients a means of control over who can access their information.

In terms of using patients' information for research purposes, the interactive nature of the proposed architecture eliminates the possibility of successful privacy attacks that could be performed on the dataset when it is released for research purposes. The recursive l -diversity together with controlling the number of tuples and percentage of tuples that contain sensitive target values eliminates the ability of adversaries to identify an individual patient or associate a certain value to individuals in any released dataset.

Having discussed the privacy-preserving strategies and techniques followed in the proposed cloud architectural design, the following chapter presents a demonstration of how these strategies and techniques are incorporated in the implementation of the proposed architecture. The demonstration is presented in a scenario-based instantiation that is explained in more detail.

Chapter 6: System Instantiation

Having discussed the proposed architectural design in terms of its architectural components, mechanisms, and strategies followed for sharing healthcare information in a privacy-preserving manner, this chapter presents a scenario-based instantiation of the proposed architectural design. The instantiation aims to exemplify how the proposed architectural design enables for sharing healthcare information in a privacy-preserving manner. The instantiation covers three main aspects: storing patients' information on the cloud, accessing patients' information while it is stored in the cloud, and finally facilitating the use of patients' information for research purposes without questioning the privacy of individual patients.

The following section presents an example scenario of how patient information is stored in the cloud and accessed by disparate practitioners when required. Section 2 presents another scenario-based instantiation of using patient information for research purposes without violating the privacy of individual patients. Finally, the last section presents the discussion and conclusion of the chapter.

The main objective of the proposed architecture design is to store healthcare information on the cloud and access it for genuine purposes in a privacy-preserving manner. This section presents a scenario-based instantiation to demonstrate how the proposed architecture design meets this objective. The instantiation is presented in two parts: the first part presents a scenario in which a patient enrolls in the system and their information is stored on the cloud, while the second part presents a scenario of accessing the information stored on the cloud for medical treatment purposes. Both parts of the instantiation are described in a step-by-step fashion.

6.1 Storing patient information on the cloud - Scenario

The scenario involves a patient (Bob) who wishes to enroll in the system and store his health-related information on the cloud. Bob visits a doctor and requests to enroll in the system. The assumption in this scenario is that the doctor has access to (doc-1, doc-2, and doc-3) of patients' information and has been given Bob's consent. The process of storing Bob's information on the cloud comprises 3 main stages namely: information preparation, authentication, authorization, and finally information storage.

Upon Bob's request, the information related to Bob is entered into the system by the doctor via the UA interface. The doctor clicks on the **ENROL** button on the application interface so that

information is stored in the cloud by forwarding it to the RA. As part of Bob's information, Bob has the option of granting permanent consent to particular practitioners to access his information whenever required. For example, he may wish to authorize his family doctor to access his information. This piece of information will be stored on the RA for user authorization purposes which are discussed further in this chapter.

Stage 1: Information Preparation (Searchable Symmetric Encryption)

Prior to forwarding the information to the RA, Bob's information undergoes a number of pre-processing algorithmic operations performed by UA as preparation for storage. The operations aim to encrypt Bob's information for storing it on the cloud in a searchable manner. The information preparation process happens in 5 steps as presented in figure (6.1).

Step 1: The first step in the preparation process is generating a secret root key (S_{KR}) for the patient (Bob). For this, the UA employs the KeyGen algorithm to generate a random S_{KR} . This S_{KR} is used for encrypting the trapdoors (explained further in this section) which are required to decrypt patients' documents. Every patient has a unique S_{KR} .

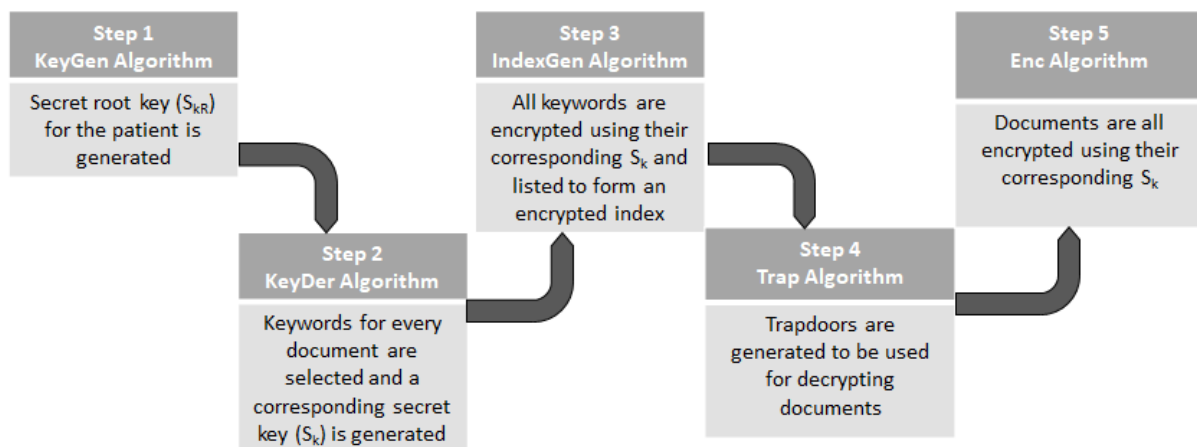


Figure (6.1) Information pre-processing steps performed by the user application

Step 2: The second step involves selecting several keywords from each document and generating a secret key (S_k) to encrypt them. The generated S_k for encrypting the selected keywords will be used to encrypt the entire document which contains these keywords. A unique secret key S_k will be generated for every document. For this, the UA employs a KeyDer algorithm that selects keywords and generates (S_k). For exemplifying purposes, the keywords selected for patient documents are predetermined as the following:

Document 1	Document 2	Document 3
Doc-1 Phone Number Address	Doc-2 Next Kin History	Doc-3 Blood Type Mental Health

Step 3: For every document, the selected keywords are encrypted using their corresponding S_k . The goal in this step is to create an encrypted index for each document to identify it while encrypted. For this, the UA employs an algorithm that takes in keywords and S_k as input and outputs keywords ciphertexts. Below is an example of encrypted indexes using the keywords selected in the previous step.

Doc-1 Index	Doc-2 Index	Doc-3 Index
B&M8\$BHY OIW0ESK(J NV#@NJSA	B&M8\$POL IK&HN\$5W LB#W@MZ	BB&M8\$9IK OD^J*~@HNY OJLS&&FDG%9

Step 4: After encrypted indexes are generated for all documents, the UA employs an algorithm that groups the ciphertexts of keywords with their corresponding S_k for each document to create

Doc-1 Trapdoor	Doc-2 Trapdoor	Doc-3 Trapdoor
Sk B&M8\$BHY OIW0ESK(J NV#@NJSA	Sk B&M8\$POL IK&HN\$5W LB#W@MZ	Sk BB&M8\$9IK OD^J*~@HNY OJLS&&FDG%9

trapdoors for documents. These trapdoors will be used to decrypt the documents. Below are the trapdoors created for the three encrypted indexes created in the previous step.

Step 5: The last step in the information preparation process involves encrypting the patient's documents and their corresponding trapdoors. Each document is encrypted using the S_k that is included in its corresponding trapdoor, and trapdoors are encrypted using the S_{kR} which was generated in the first step.

Note: In a production implementation of the system, updating indexes or secret keys can happen whenever needed by a user who is authorized to update such information.

Stage 2: Authentication and Authorization

When Bob's information is pre-processed, it is forwarded to the RA in a form of Request of Enrol (ROE). The ROE includes three sections as illustrated in figure (6.2). The first section includes information that is required to identify the user (doctor), the patient (Bob), users who are granted consent (by Bob) to access Bob's information, and finally method of obtaining Bob's consent for users to access his information. The second section includes the information

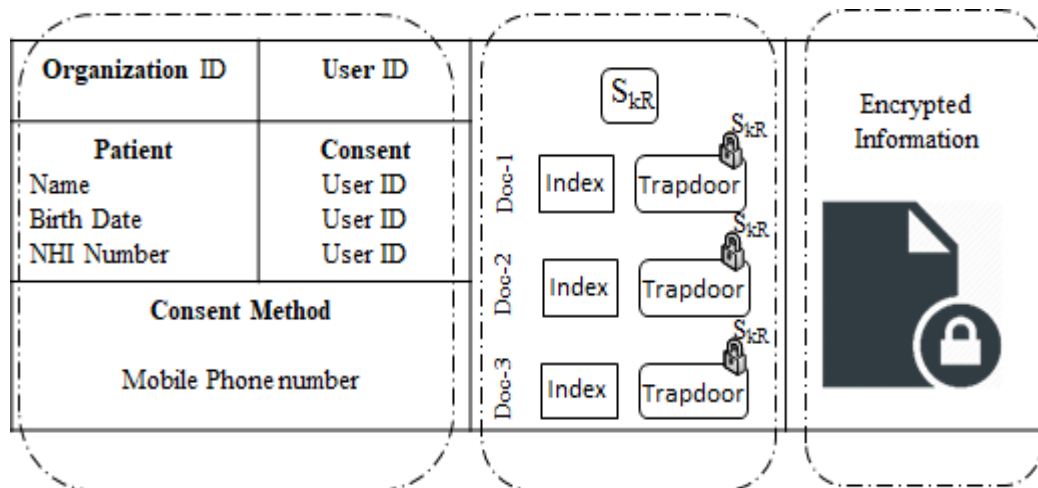


Figure (6.2) ROE to store Bob's information

required to identify and decrypt Bob's information, while the encrypted Bob's information (3 documents) is included in the third section.

When the RA receives the request from the doctor, it authenticates the doctor and forwards the user's information (contained in the first section of the ROE) and Bob's phone number to the CSR for authorization. The CSR then sends a text message to Bob requesting consent to store his information on the cloud. The content of the message includes:

Please reply **YES** to authorize **(doctor name)** from at **(organization name)** to enroll you and store your health information on the system.

Upon receiving a YES reply from Bob, the CSR sends a confirmation of authorization to the RA.

Stage 3: Information Storage

When the RA receives confirmation from the CSR that the user is authorized to store Bob's information, it does the following actions:

1. It generates a unique code for the patient referred to as (System ID).
2. Sends Bob's encrypted information labeled by Bob's system ID to the CSP
3. Sends the encrypted trapdoors to the SKA for storage. Information sent to the SKA is also labeled by Bob's system ID.
4. The information sent to both CSP and SKA is deleted from the RA.
5. The RA stores the following information:
 - a. Bob's identification information
 - b. Bob's system ID,
 - c. Bob's S_{KR} ,
 - d. Document indexes
 - e. Names of users who have permanent consent to access Bob's information (if Bob has provided any).
 - f. Information required for obtaining Bob's temporary consent

The process of enrolling Bob in the system and storing his information on the cloud results in having Bob's information stored in the cloud in encrypted form. The decryption of Bob's information can only happen using the secret keys that are stored on the SKA. Below is the state-of-the-art of Bob's information while stored in the cloud:

1. Bob's information is stored in encrypted form and labeled by Bob's system-generated ID. The cloud provider is not able to learn the content of the information.
2. The trapdoors are encrypted using Bob's S_{KR} and stored on the SKA labeled by Bob's system ID. The SKA is unable to learn the content of the trapdoors without having Bob's S_{KR} that is stored on the RA.
3. The RA is the only entity in the system that can identify Bob in the system and his S_{KR} . The RA stores all the information that is required to access Bob's information as presented in figure (6.3).

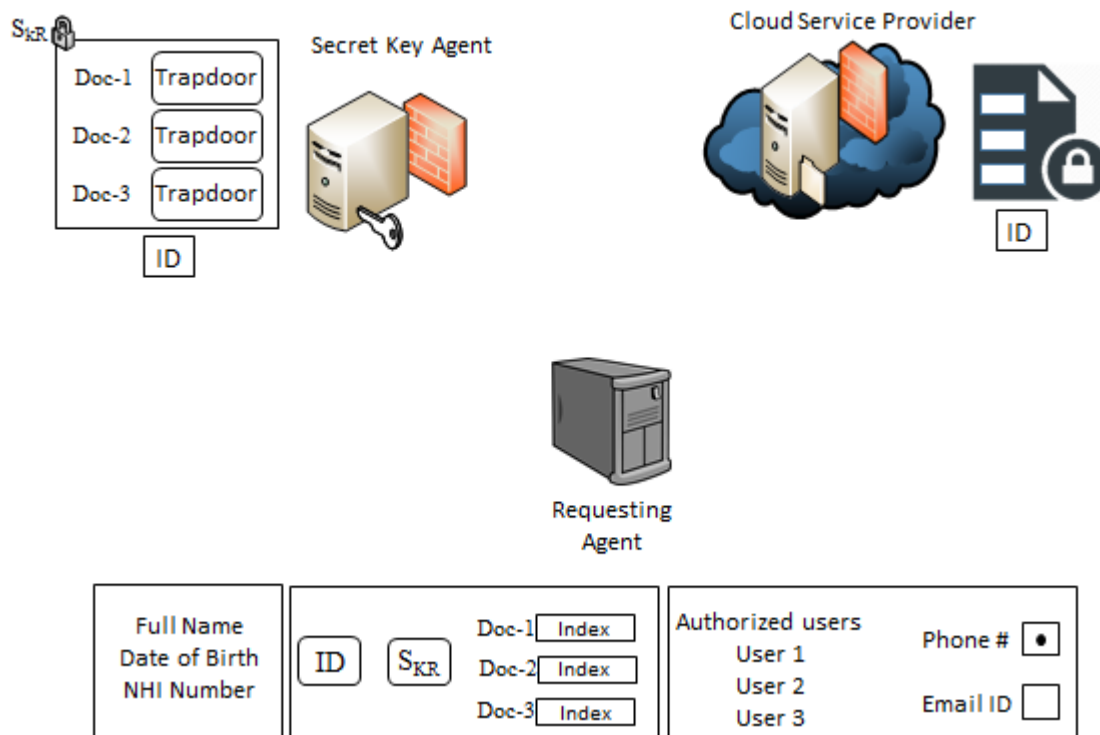


Figure (6.3) Bob's information stored in the proposed architecture

Therefore, accessing Bob's information can only happen through collaborative interactions among CSP, SKA, and the RA. Compromising 1 or 2 of these architectural components will be fruitless to any disparate party in terms of accessing Bob's information.

Having discussed the process of storing Bob's information in the cloud, the following subsection presents the process of accessing Bob's information for genuine reasons such as providing healthcare to a patient. For this, the scenario presented in the following subsection involves the same patient (Bob) requiring healthcare assistance by a different medical practitioner who also has access to the system.

6.2 Accessing stored patient information – Scenario

Bob -after a few weeks- visits a hospital and requires urgent treatment. He walks into the emergency department and meets one of the nurses in charge. The nurse requires accessing Bob's information for urgent medical treatment and updating his records to include information about Bob's visit, medical condition, and other information related to his visit.

6.2.1 Protocol to access information stored in the cloud

The process of accessing Bob's information comprises 4 stages as the following:

Stage 1: Generating user request

The user (nurse) enters Bob's basic information into her system application and clicks on the **REQUEST** button to generate a user request. The user request includes information about both Bob and the nurse. By clicking on the **REQUEST** button, a request is generated and forwarded to the RA.

Organization	ID
User	ID
Patient	Name Birth Date NHI Number

Stage 2: Authentication and Authorization

When the RA receives the request from the user (nurse), it authenticates the user and forwards the request to the CSR for authorization. For this, the RA does the following actions:

1. It searches for Bob's information using his basic information and finds his System ID.
2. It sends a request of authorization to the CSR. The request includes the following information:
 - a. Information that is required to identify the nurse which includes organization ID and user ID.
 - b. List of users who have permanent consent to access Bob's information.
 - c. Bob's mobile number for requesting his consent if the nurse is not issued with permanent consent to access Bob's information.

Organization ID	Nurse ID
Authorized Users	Bob's Phone number
User ID User ID User ID User ID	

When the CSR receives the request from the RA, it does the following actions:

1. It searches for the nurse information to identify her access rights to patient information. This happens by searching through the list of users that is stored locally on the CSR.
2. It checks if the nurse is permanently consented to access Bob's information using the list of users who have permanent consent to access Bob's information.

The CSR finds out that the nurse is allowed access doc-1 and doc-2 (Em_V) of patients' information, but she is not permanently consented to access Bob's information, therefore, Bob's consent is required.

3. The SCR sends a request of consent to Bob in the form of a text message. The content of the message includes:

Please reply **YES** to temporarily authorize (**nurse name**) at (**organization name**) to access your health information.

4. Upon receiving a YES from Bob, the nurse becomes temporarily authorized to access Bob's information. The CSR sends a confirmation of authorization to the RA. The confirmation of authorization includes the information category that the nurse can access (doc-1 and doc-2) and confirmation of obtaining Bob's consent to access his information. The nurse is then added temporarily to the list of authorized users (stored on the RA) as a temporarily authorized user. However, any authorization granted by the CSR remains valid for 1 hour, after that it is automatically deleted from the list of authorized users.

Note: An assumption in this instantiation is that Bob is conscious and able to provide consent for accessing his information, however, Bob's authentication could happen via a biometric method such as fingerprint (in case of emergency). The main goal of this instantiation is to elaborate on how information is released in a privacy-preserving manner. Further configurations may be considered in the actual implementation of the system design.

Stage 3: Releasing Information

Upon receiving confirmation of authorization from the SCR, the RA forwards requests to both, the CSP and the SKA to send Bob's information to the nurse. As explained earlier, the information stored on the CSP is different from the information stored on the SKA therefore, the RA sends different requests to both of them.

As illustrated in figure (6.4), the request to the CSP includes the following information:

- a. Bob's system ID
- b. Indexes of doc-1 and doc-2
- c. The nurse's application address

While the information included in the request to the SKA includes:

- a. Bob's system ID and S_{kR}
- b. Trapdoor-1 and Trapdoor-2 tags
- c. The nurse's application address

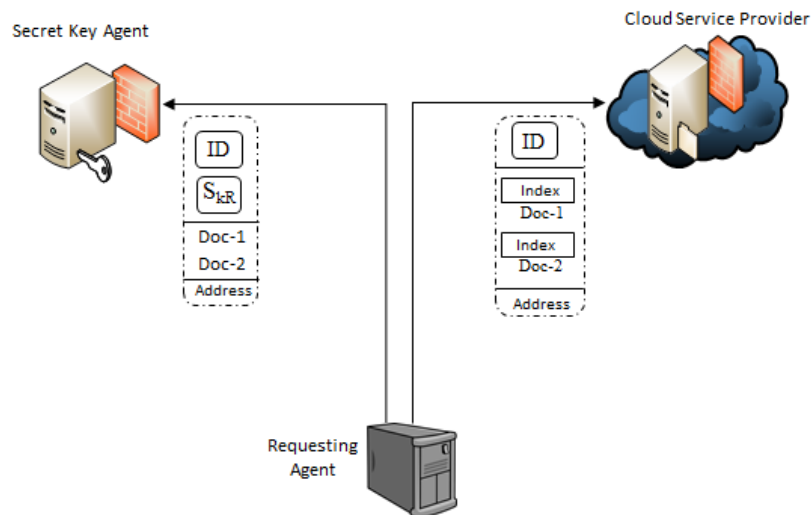


Figure (6.4) Requesting information from CSP and SKA

As presented in figure (6.5), when the RA requests are received by the CSP and the SKA, they do the following actions:

The CSP:

1. Searches for Bob's information using Bob's system ID
2. Searches for the doc-1 and doc-2 using their indexes. This happens by an algorithm employed by the CSP that uses the indexes to search for documents
3. Sends the identified documents (doc-1 and doc-2) to the nurse using her application's physical address.

The SKA:

1. Searches for the encrypted trapdoors using Bob's system ID
2. Decrypts the trapdoors using Bob's S_{kR}
3. Sends trapdoors for doc-1 and doc-2 to the nurse application using her application physical address.
4. Re-encrypts the trapdoors using the same S_{kR} and drops the S_{kR} (deletes it).

Stage 4: Decrypting Information

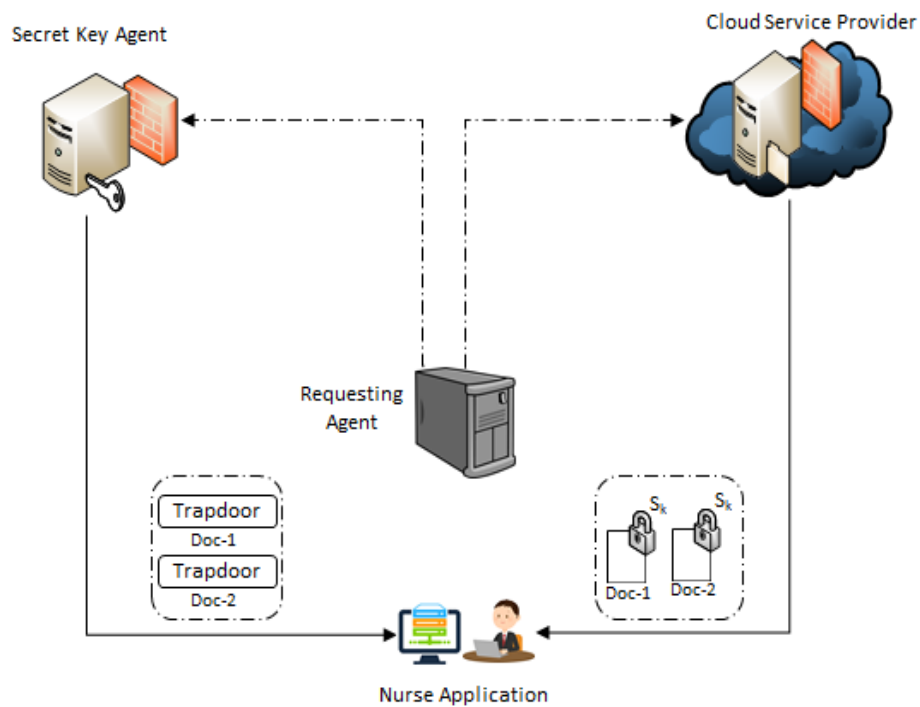


Figure (6.5) Releasing information to the nurse application

When the information from CSP and SKA is received by the nurse's application, doc-1 and doc-2 are identified and decrypted using their corresponding trapdoors. When information is decrypted, files in each document appears in their predetermined fields on the nurse's UA. Fields that belong to the files contained in doc-3 remain blank. The nurse application stores the trapdoors temporarily to be used for re-encrypting the information which is further explained in the following section.

6.2.2 Updating patient information

Assuming that the nurse has made an update on Bob's information such as information related to current medication. The nurse clicks on the **UPDATE** button on her UA interface.

Stage 1: The nurse's UA encrypts doc-1 and doc-2 using their secret keys obtained from the trapdoors. The encrypted information (doc-1 and doc-2) is forwarded to the RA.

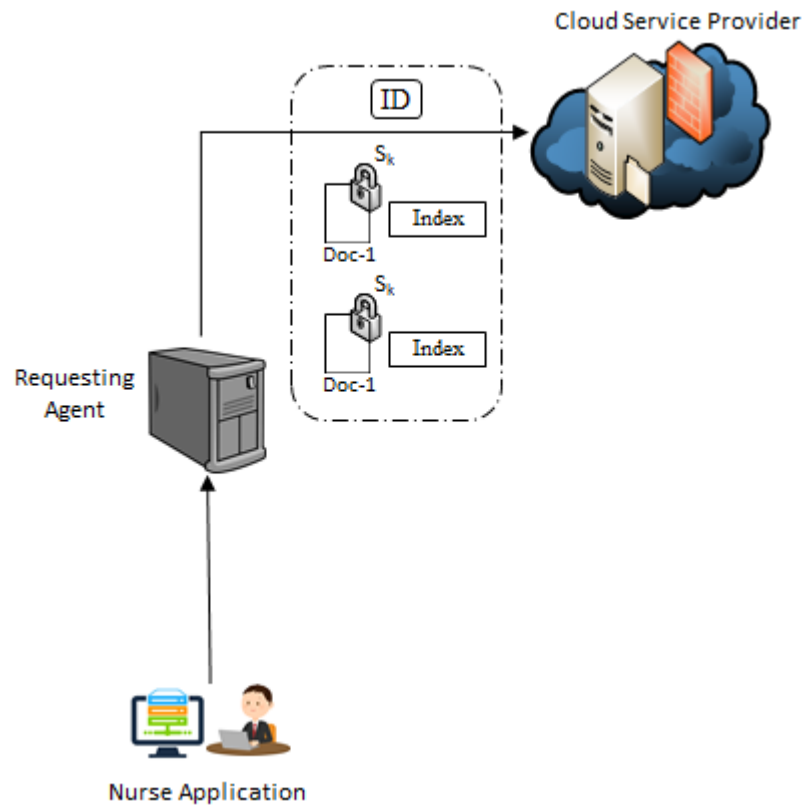


Figure (6.6) Updating patient information

Stage 2: The RA receives the request from the nurse and does the following:

1. It searches for Bob's information to identify him.
2. It searches through the list of authorized users to access Bob's information and finds the nurse listed as temporarily authorized users to access doc-1 and doc-2 of patients' information.
3. It forwards the encrypted information, indexes for doc-1 and doc-2, and Bob's system ID to the CSP.

Stage 3: When the CSP receives the information from the RA, it does the following actions:

1. It searches for Bob's encrypted documents using indexes and system ID.

2. It identifies the documents using the indexes and replaces them by the new ones
3. It deletes the indexes received from the RA.

6.2.3 Security Analysis of the Proposed system design

Although the proposed system design focuses on preserving the privacy of information when storing it on the cloud and sharing it among different parties in the healthcare domain, the security of this information is also preserved. The proposed system design incorporates several security measures on different levels of the architecture on both, client-side and cloud side.

- **Client-side security / User Application Level**

Having a standard user application that is required to access the system provides a line of security in the proposed system. The user application in the real implementation of the system design allows only for certain operations to be performed by the user. Users in the proposed system design are given accessibility to the system that is controlled by the enabled features/functions of the user application. For example, a nurse's log-in credentials enable certain functions on the user application to access the system, meaning that a nurse cannot perform operations to modify the way information is stored on the system. In the real implementation of the system, the login credentials can be a pair of user ID and Password, or through the Bio-metric system.

Moreover, the encryption and decryption processes are not controlled by the user. The user application is responsible to perform all these operations internally without the ability of the user to understand how these operations are performed. The user application in the real implementation of the system may have a characteristic to hide all information that is related to the information encryption or decryption of information. For example, when the secret key is received by the user application, the user should not be able to read it or access it at any stage, it remains hidden and only used internally by the system to re-encrypt the information prior to storing it. The user application internally processes all information as part of the system as described earlier in Section 6.1.

- **Cloud-side security / Access control**

The proposed system design employs the concept of user identity management (U-IDM) which is the Cloud Service Registry (CSR). The CSR is a component that is not located at the client-side, meaning that users cannot attempt to add or modify access rights to the system. Moreover, accessing the system can happen through requests which are sent from the user application to

the Requesting Agent (RA). The RA -as described earlier- works as the gateman who must authenticate and authorize users before their requests are processed further. The collaboration between the RA and the CSR is the only way to forward users' requests to access the system. Therefore, there are three security stations in the system that the user must go through to access the system as presented in figure (6.7).

The first station is at the client side which involves entering the users' credentials to access the system. Users' credentials in the real implementation can be through information entered by the system (user ID and Password) or other forms of credentials such as figure prints. The second station is at the cloud side which involves authenticating the user. The user must be authenticated by the RA component before the authorization process takes place. Once a user is authenticated by the RA, the authorization process happens by the CSR. The CSR -as mentioned earlier- is not accessible by the user, it only communicates with the RA component.

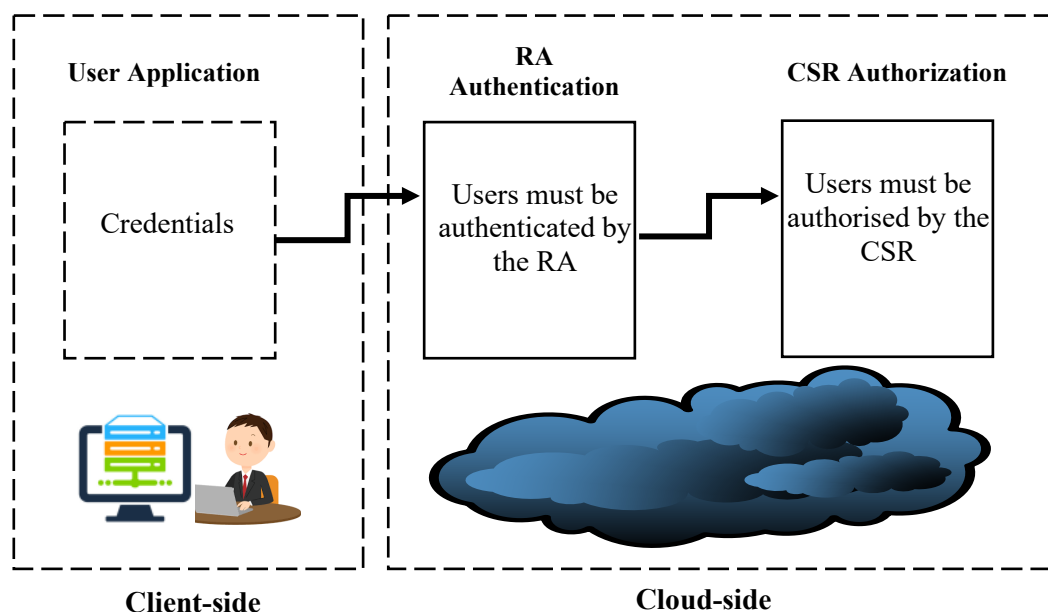


Figure (6.7) Security stations for accessing the system

- The separation of information

The separation of information, while it is stored in the system, makes it difficult to perform any unauthorized actions that could lead to reading the information that is stored on the cloud, especially because the cloud provider cannot learn the content of the information that is stored on the cloud. The proposed system design provides means of security to the information in its simplest implementation due to: (1) storing encrypted information and decryption keys on

different components of the architecture, (2) requiring the collaboration of three different components of the architecture to gain access to the information that is stored in the system. The information required to identify patients and their information is stored on the RA. Therefore, compromising any of these three components (CSP, SKA, or RA) will be fruitless to any party in terms of reading the information stored on the cloud.

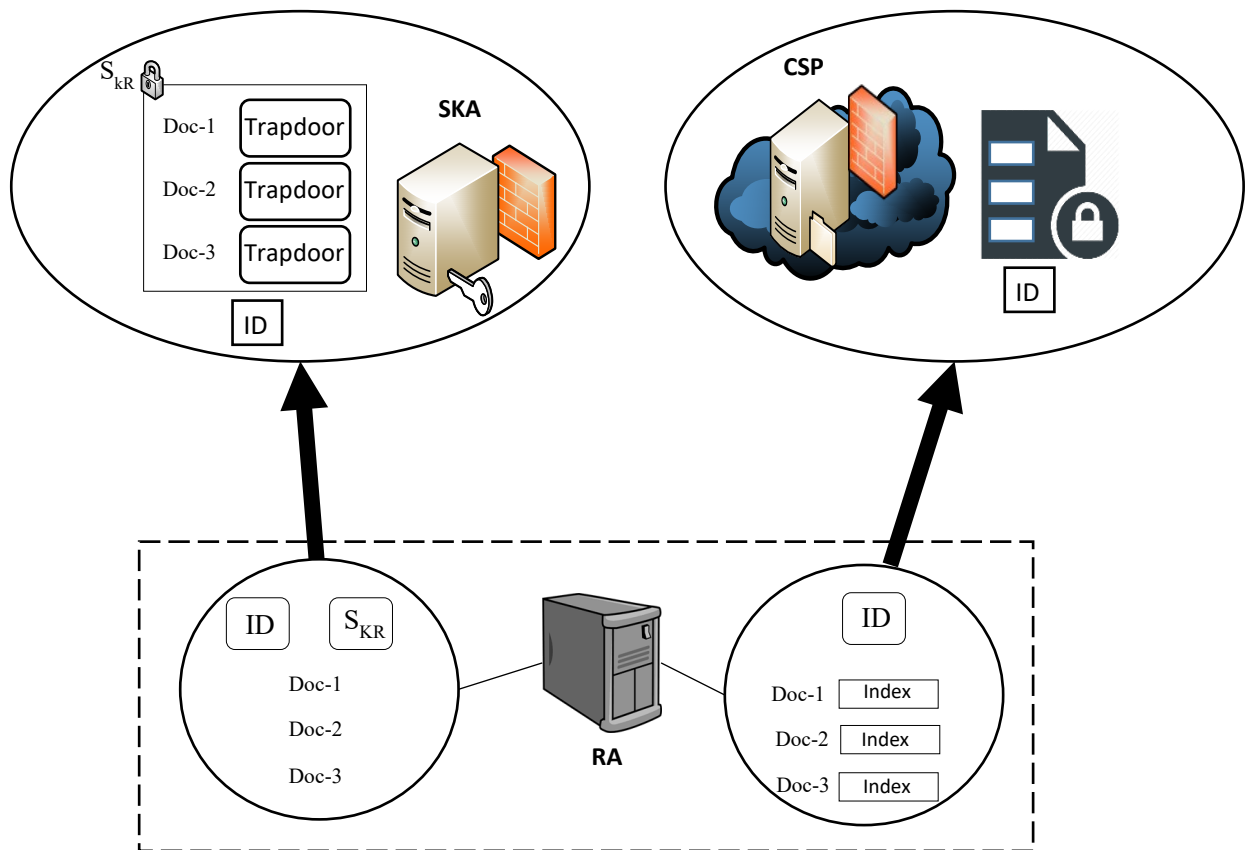


Figure (6.8) Separation of information stored in the system

As seen in Figure (6.8), information is stored in three different components of the system, and accessing it requires the collaboration of these components. To access information about a particular patient, the RA needs to send the patient ID and the documents' tags to the CSP in order to identify the right information of the patient. The RA needs to send the patient ID, S_{KR} , and the Trapdoors tags in order to identify the patient, decrypt the trapdoors, and release the right trapdoors to decrypt the patient's information. Therefore, without this collaboration of components, accessing information is not possible.

- Communication channels

Another security measure that is incorporated in the proposed system design is the communication channels within the system. Since the collaboration among the architecture components is the only way to access the information that is stored in the system, the communication channels among these components also add another measure of security to the entire system as seen in Figure (6.9).

The absence of communication between the CSP and the SKA maintains the separation of information that is stored on both of them. The CSP is not able to learn the content of the information stored on it, neither obtain the information required to decrypt it (secret keys).

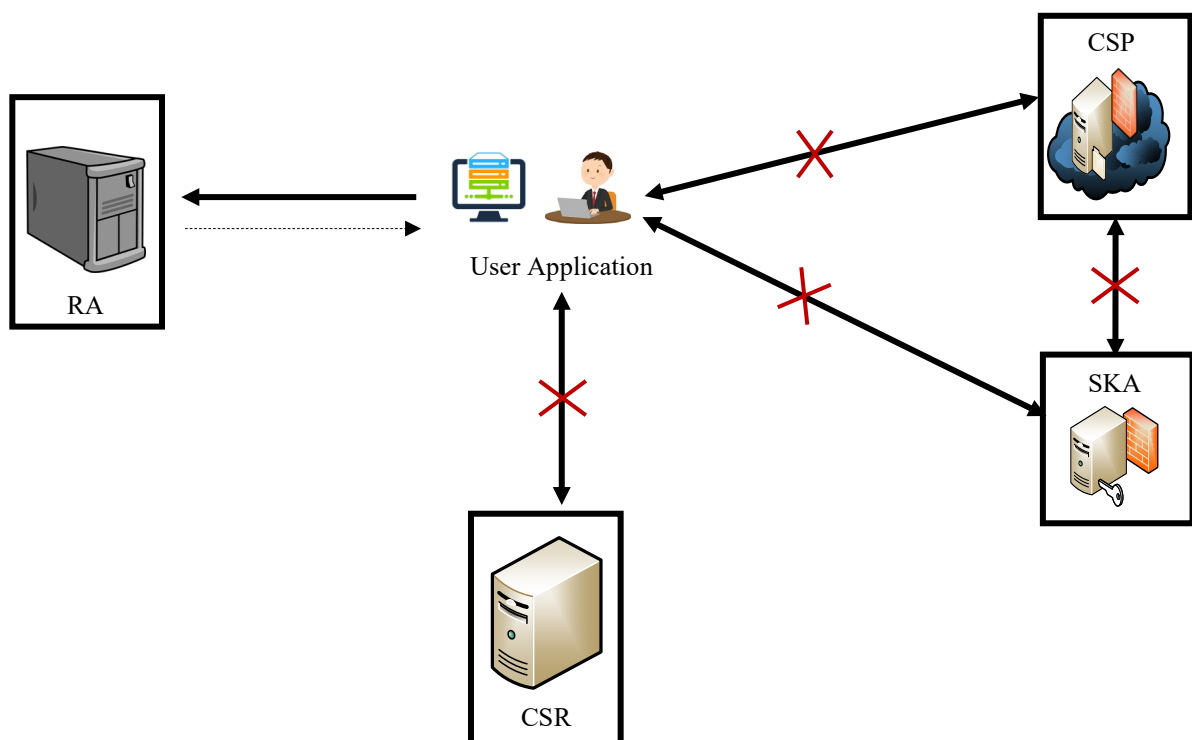


Figure (6.9) Communication channels as a measure of security

The absence of direct communication between the user and both CSP and SKA makes it difficult to perform any attempt to directly access the information that is stored on them. Access to the system can only happen through the RA, and this eliminates the chance of any attempt to break into the information that is stored in the system.

The one-way communication between the RA and the user also eliminates the chance of performing actions to break into the information that is stored in the system. The one-way communication as explained earlier in Section (5.1.3) aims to limit the responses that the user

receives from the RA, which makes the user in a position to only request information without much interactivity with the RA.

Finally, the absence of information between the user and the CSR makes it difficult for any user to purposefully modify access rights or authorization privileges on the CSR without going through an authenticated process. Access to the CSR can only happen with the RA, and any modification on the list of users/privileges may only happen through the RA.

The security of the proposed system can always be enhanced by other measures that could be incorporated into the system in its real implementation. For example, employing an advanced searchable encryption scheme, activity log, other means of physical security, advanced methods of obtaining login credentials from users.

6.3 Architecture Implementation

Having discussed the components and protocol of the proposed architecture, the architecture is further implemented and adapted to data sharing use. The authors in (Nunamaker, Chen, & Purdin, 1991) state that when the proposed solution of the research problem cannot be proven mathematically and tested empirically, or if it is a proposal of a new way of doing things, researchers may elect to develop a system to demonstrate its validity as a solution, based on the suggested new methods, techniques, or design. Therefore, the researcher has implemented the designed cloud architecture for validating the concept of the proposed architecture.

Amazon Web Services (AWS)

The proposed architecture was built using Amazon Web Services (AWS) which provides cost-effective cloud computing solutions (Amazon, 2019). AWS Software Development Kit (SDK) was used with Java language to implement and test the proposed architecture design and validate its concept. The implementation diagram of the proposed system is shown in Appendix D.

AWS is a comprehensive and evolving cloud computing platform provided by Amazon. It includes a mixture of infrastructure as a service (IaaS), platform as a service (PaaS), and packaged software as a service (SaaS) offerings (Rouse, Amazon Web Services (AWS), 2019). It is a subsidiary of Amazon that provides on-demand cloud computing platforms and application program interfaces (APIs) to individuals, companies, and governments, on a metered pay-as-you-go basis. These services are cloud computing services that provide a set of

primitive abstract technical infrastructure and distributed computing building blocks and tools. Such computing blocks and tools can be used to create and deploy any type of cloud application therefore, it was found suitable to use for this research.

6.3.1 AWS services used

There is a wide range of different business purposes global cloud-based products offered by AWS which include storage, compute, databases, analytics, networking, mobile, developer tools, management tools, IoT, security, and enterprise application. This section presents the AWS services that were used in this research for implementing and validating the proposed cloud architecture. All information about AWS services presented in this section is taken from the documentation of the AWS (AWS, 2019).

Elastic Compute Cloud (EC2)

Amazon Elastic Compute Cloud (EC2) is a service that provides a scalable computing capacity in the Amazon Web Services (AWS) cloud. It is a virtual machine in the cloud which is controlled on the operating system level. EC2 service enables to launch as many virtual servers as needed, configure security and networking, and manage storage. It is a web service that provides secure, resizable compute capacity in the cloud. EC2 service was used in the implementation of the proposed architecture as the Requesting Agent (RA). The RA component plays a significant role in the proposed architecture; it acts as a gateman who allows only authenticated and authorized users to access the information stored within the system.

Virtual Private Cloud (VPC)

Virtual Private Cloud (VPC) is a service that enables for establishing a secure and private tunnel from the researcher device to the AWS global network. VPC service is comprised of two services namely Site-to-Site VPC and Client VPC. The Site-to-Site VPC service enables for securely connecting on-premises network or branch office to a privately owned virtual private cloud, while the client VPC service enables to securely connect users to AWS or on-premises networks. The VPC service enabled the researcher to launch AWS resources into a private network. The goal of using this service was to create a virtual cloud network that closely resembles a traditional network that could be run in the researcher's own data center with the benefits of the scalable infrastructure of AWS.

Amazon Machine Image (AMI) and Instances

Amazon Machine Image (AMI) is a special type of virtual appliance that is used to create a virtual machine within the Amazon Elastic Compute Cloud (EC2). It serves as the basic unit of deployment for services delivered using EC2. AMI is a template that contains a software configuration such as an operating system, an application server, and applications. From an AMI, an instance is launched, which is a copy of the AMI running as a virtual server in the cloud. The instance is a virtual server in the cloud. The configurations of the instance are a copy of the AMI that is specified when the instance is launched.

Identity and Access Management (IAM)

Identity and Access Management (IAM) is a security service that aims to manage users, assign policies, form groups to manage multiple users. It helps to securely manage access to cloud services and resources. It uses permissions to allow and deny access to AWS resources. The IAM service was used to enforce policies related to authentication and authorization. Users in the proposed system design are allowed to access patients' documents based on their roles in the healthcare domain, therefore, the IAM service enabled the control of access to patients' information.

Simple Storage Service (S3)

Amazon Simple Storage Service (S3) is an object storage service. S3 service has a simple web services interface that users can use to store and retrieve information, at any time, from anywhere on the web. It is virtual limitless storage on the internet. More importantly, S3 service provides easy-to-use management features so users can organize their data and configure finely-tuned access control to meet specific compliance requirements. S3 service was used for the implementation of the proposed architecture to store and access information.

Key Management Service (KMS)

Key Management Service (KMS) is a service that allows users to create, delete, and control keys for encrypting and decrypting information that is stored in AWS databases and products. KMS service enables to easily manage encryption keys and control the use of encryption across a wide range of AWS services within the application. KMS service was used in the implementation of the proposed architecture to manage the encryption keys in terms of generating, storing, and releasing them to users upon their requests to access patients' information.

Relational Database Service (RDS)

Amazon Relational Database Service is a service that enables to set up, operate, and scale a relational database in the cloud. RDS is available on several database instance types and provides six familiar database engines to choose from which SQL is one of them. Therefore, RDS was used in the implementation for the execution of queries and transactions on the stored information including patients' information (documents), users, and trapdoors.

6.3.2 Implementation Objectives

The implementation of the proposed architecture aimed to elaborate on how the proposed architecture enables for collaborative use of patients' information in a privacy-preserving manner. The main objective of the implementation was twofold: firstly, to elaborate on how patients' information can be collaboratively shared and used in the proposed architecture with assurance to its privacy protection, and secondly to illustrate on how patients' information is protected from a number of privacy-related threats including confidentiality and unauthorized access.

The elaboration is presented in two parts: the first part presents a scenario that involves a patient who walks into a hospital for urgent medical treatment and is seen by a nurse. The goal of this part is to show how a user (nurse) can access a patient's information according to certain access rights without questioning the privacy of the information. The elaboration also aims to validate the concept of information separation in real cloud-based application contexts.

The second part of the elaboration presents the results of tests that have been performed on the implemented architecture. The architecture was tested in terms of its ability to preserve the privacy of information while it is stored on the cloud. Four tests were performed which covered the following aspects:

1. The ability of the cloud provider to access and read the information that is stored in the cloud
2. The ability of unauthorized users to access patients' information stored in the cloud

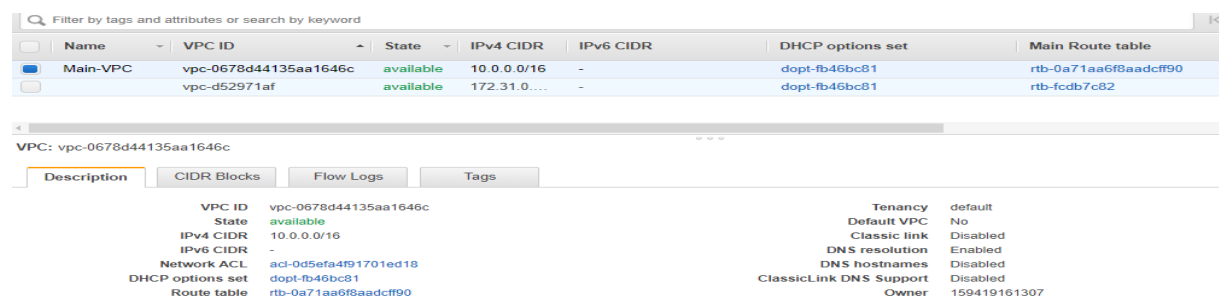
The following section presents information about the implementation of the cloud architecture in terms of its architectural components and the privacy-preserving techniques. The implementation aims to mirror the proposed architecture design in terms of the employed components and the privacy-preserving techniques for the goal of validating the concept of the proposed architectural design.

6.3.3 Implementation setup

This section presents the main aspects of the implemented architecture in terms of its characteristics and the privacy-preserving measures incorporated. The implementation diagram is annexed in Appendix D.

Virtual Private Cloud

For the implementation of the proposed architecture, a virtual private cloud (VPC) was created. The VPC represents the entire proposed system architecture and it is completely isolated from the internet as seen in the figure below.



The screenshot shows the AWS VPC console. At the top, there's a table listing VPCs. Below that, the details for VPC 'vpc-0678d44135aa1646c' are shown. The 'Description' tab is selected, displaying various attributes.

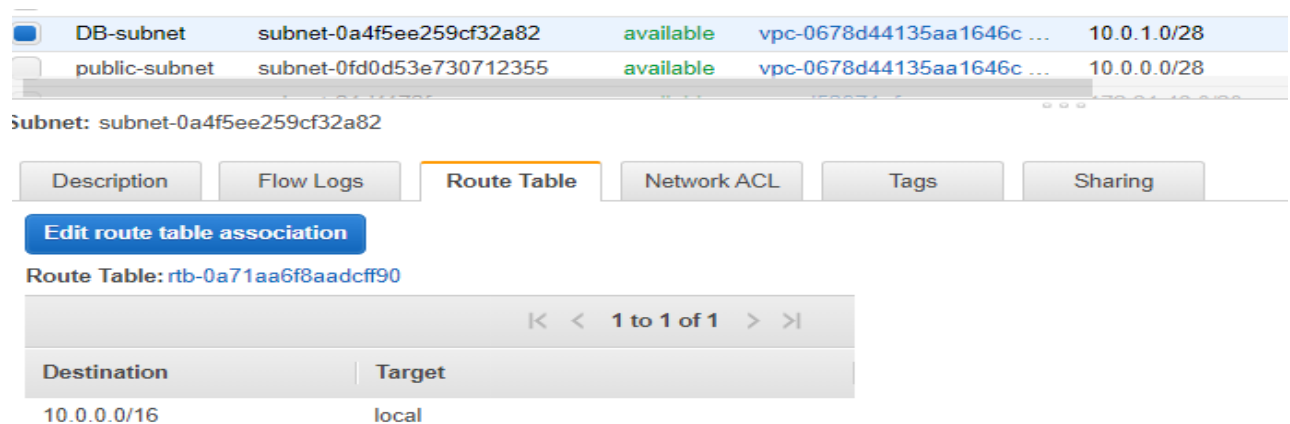
Name	VPC ID	State	IPv4 CIDR	IPv6 CIDR	DHCP options set	Main Route table
Main-VPC	vpc-0678d44135aa1646c	available	10.0.0.0/16	-	dopt-fb46bc81	rtb-0a71aa6f8aadcf90
	vpc-d52971af	available	172.31.0.0/16	-	dopt-fb46bc81	rtb-fc8b7c82

Attribute	Value	Attribute	Value
VPC ID	vpc-0678d44135aa1646c	Tenancy	default
State	available	Default VPC	No
IPv4 CIDR	10.0.0.0/16	Classic link	Disabled
IPv6 CIDR	-	DNS resolution	Enabled
Network ACL	acl-0d5efa4f91701ed18	DNS hostnames	Disabled
DHCP options set	dopt-fb46bc81	ClassicLink DNS Support	Disabled
Route table	rtb-0a71aa6f8aadcf90	Owner	159419161307

Figure (6.10) isolated virtual private cloud

Private and Public Subnets

The VPC has two subnets: the first subnet does not allow access from the internet; it is only accessed locally. The second subnet enables internet access to allow for communication with the user application.



The screenshot shows the AWS VPC console. At the top, there's a table listing subnets. Below that, the details for subnet 'subnet-0a4f5ee259cf32a82' are shown. The 'Route Table' tab is selected, displaying the route table association.

Name	Subnet ID	State	VPC ID	IPv4 CIDR
DB-subnet	subnet-0a4f5ee259cf32a82	available	vpc-0678d44135aa1646c	10.0.1.0/28
public-subnet	subnet-0fd0d53e730712355	available	vpc-0678d44135aa1646c	10.0.0.0/28

Destination	Target
10.0.0.0/16	local

Figure (6.11) Private and public subnets of the VPC

Database security

The public subnet has limited access and could be communicated with through the internet; however, only registered IP addresses can communicate with it. The EC2 represents the RA in the proposed design. Access to it happens only through a particular port. Moreover, the database in the implemented design only accepts SQL traffic.

The screenshot displays two AWS Security Groups. The top section shows the details for 'sg-0290255798f099a47', which is associated with the 'launch-wizard-4' VPC. It lists an inbound rule for Custom TCP Rule on port 6666 from source 156.62.3.187/32. Below this, the instance 'i-0e4e4f81af1590879' (RA in VPC) is shown with its status and various attributes like Instance ID, state, type, and network configuration.

The bottom section shows the details for 'sg-01e917a5c7d579c83', associated with the 'RDS-SG' VPC. It lists two inbound rules for MySQL/Aurora on port 3306 from source 0.0.0.0/0.

Name	Group ID	Group Name	VPC ID	Owner
sg-0290255798f099a47	launch-wizard-4	vpc-0678d44135aa1646c	159419161307	

Security Group: sg-0290255798f099a47

Type	Protocol	Port Range	Source
Custom TCP Rule	TCP	6666	156.62.3.187/32

Instance: i-0e4e4f81af1590879 (RA in VPC) Elastic IP: 54.89.89.197

Type	Protocol	Port Range	Source
MySQL/Aurora	TCP	3306	0.0.0.0/0
MySQL/Aurora	TCP	3306	:::0

Figure (6.12) Database security measures

Patients' documents and their associated trapdoors are stored in two different places that are not accessible through the internet as seen in Figure (6.13). Access to them can only happen locally. Access to patients' documents can only happen by requests from the CSR, while the trapdoors can only be accessed through the SKA.

S3 buckets [Discover the console](#)

Search for buckets All access types

+ Create bucket Edit public access settings Empty Delete 4 Buckets 1 Regions

<input type="checkbox"/> Bucket name	Access	Region	Date created
<input type="checkbox"/> docs-out	Bucket and objects not public	US East (N. Virginia)	Dec 10, 2019 10:15:28 AM GMT+1300
<input type="checkbox"/> patients-docs	Bucket and objects not public	US East (N. Virginia)	Dec 9, 2019 10:11:05 AM GMT+1300
<input type="checkbox"/> trapdoors	Bucket and objects not public	US East (N. Virginia)	Dec 10, 2019 12:36:01 AM GMT+1300
<input type="checkbox"/> trapdoors-out	Bucket and objects not public	US East (N. Virginia)	Dec 10, 2019 10:15:59 AM GMT+1300

Figure (6.13) Separation of information stored on the cloud

Documents stored on the cloud are all encrypted.

patients-docs

Overview Properties Permissions Management Access points

Search Type a prefix and press Enter to search. Press ESC to clear.

Upload + Create folder Download Actions US East (N. Virgin)

Viewing 1

<input type="checkbox"/> Name	Last modified	Size	Storage class
<input type="checkbox"/> d1O18ATX37YPW_encrypted	Dec 13, 2019 3:40:01 AM GMT+1300	77.0 B	Standard
<input type="checkbox"/> d4R9TK3XWX4S6_encrypted	Dec 13, 2019 3:40:01 AM GMT+1300	77.0 B	Standard
<input type="checkbox"/> dDDDU22RM68V1_encrypted	Dec 13, 2019 3:40:01 AM GMT+1300	77.0 B	Standard

Figure (6.14) Encrypted information that is stored on the cloud

Trapdoors are all encrypted and stored. Encryption of trapdoors happens inside the cloud using the KMS. Accessing them requires getting the secret root key.

trapdoors

Overview Properties Permissions Management Access points

Versioning Keep multiple versions of an object in the same bucket. Learn more <input type="radio"/> Disabled	Server access logging Set up access log records that provide details about access requests. Learn more <input type="radio"/> Disabled	Static website hosting Host a static website, which does not require server-side technologies. Learn more <input type="radio"/> Disabled	Object-level logging Record object-level API activity using the CloudTrail data events feature (additional costs). Learn more <input type="radio"/> Disabled	Default encryption Automatically encrypt objects when stored in Amazon S3. Learn more <input checked="" type="radio"/> AWS-KMS
--	---	--	--	--

Figure (6.15) AWS-KMS used for Trapdoors

Figure (6.16) presents an encrypted document of a patient's information in the implemented system and a decrypted version of it.

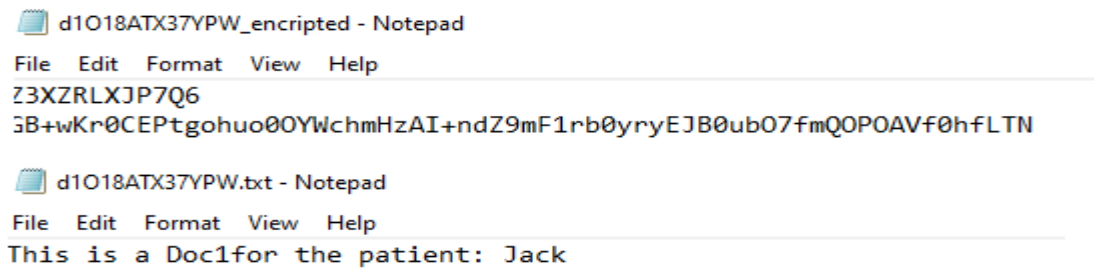


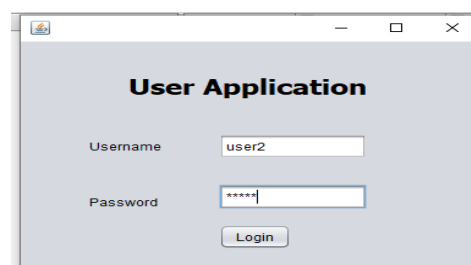
Figure (6.16) Encrypted document that is stored on the cloud

Accessing Patient information – Scenario

For the goal of testing the architecture, dummy information about 3 patients was used. Each patient had three documents in the system. Documents were all encrypted and stored in the cloud database. Encrypted documents were stored on one database (CSP) and their trapdoors were stored on a different database (SKA). The illustration involved a patient (Bob) who walked into the hospital for urgent medical treatment. The nurse wishes to access Bob's information to update information regarding Bob's visit and current medication.

User authentication

The nurse needs to login to the system for authentication purposes. The user is required to have username and password which are entered through the standard user application. Access to the system only happens through the user application. When the RA receives the user credentials (username and password), it searches for the user information in the list of registered users. The RA in the implemented architecture had a database that contains names of registered users, this database was used for authenticating users. When the user is found, authentication is confirmed.



User authorization

Figure (6.17) Login Page for authentication

When the user is authenticated by the RA, another window pops up on the user application for entering Bob's basic information as seen in Figure (6.18). Bob's information is used by the RA to identify him in the system.

Figure (6.18) Entering Patient's information

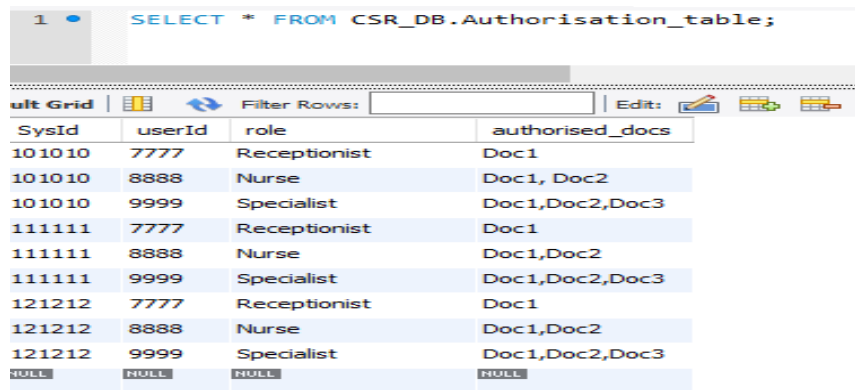
When information is entered by the nurse and forwarded to the RA, the RA searches for the patient in the list of the registered patients in the system. The RA in the implemented system had another database that contains information about all patients enrolled in the system. When the patient is found, the CSR is called for authorization. The RA sends Bob's information to the CSR along with the users' information. The following code was used in the implementation of the authorization process.

```
//Calling Cloud Service Registry

CloudServiceRegistry csr = new CloudServiceRegistry();
String authorisedDocs=csr.getAuthorisedDocumentList(userId, SysId);
System.out.println(authorisedDocs);
String [] outputDocIndex= new String[3];
String [] outputKeyIndex= new String [3];
String[] docsList=null;
String ClientReference="";
int numberOfAuthorisedDocs=0;
if(authorisedDocs==null || authorisedDocs.equals("None"))
{
    dout.writeUTF("No Access");
    dout.flush();
}
else {
    docsList = authorisedDocs.split(",");
    . . . . .
}
```

Figure (6.19) Script to call the CSR for user authorization

The CSR searches for the user in the list and finds out that the user is a nurse and is allowed to access Doc-1 and Doc-2 of patients' information. Figure (6.20) is a screenshot of the users' table that is used by the CSR to authorize users.



```
SELECT * FROM CSR_DB.Authorisation_table;
```

SysId	userId	role	authorised_docs
101010	7777	Receptionist	Doc1
101010	8888	Nurse	Doc1, Doc2
101010	9999	Specialist	Doc1,Doc2,Doc3
111111	7777	Receptionist	Doc1
111111	8888	Nurse	Doc1,Doc2
111111	9999	Specialist	Doc1,Doc2,Doc3
121212	7777	Receptionist	Doc1
121212	8888	Nurse	Doc1,Doc2
121212	9999	Specialist	Doc1,Doc2,Doc3
NULL	NULL	NULL	NULL

Figure (6.20) CSR List of users and their access privileges

The CSR confirms to the RA that the user is allowed to access Doc-1 and Doc-2 of Bob's information. The assumption made here was that Bob's has received a text message from the CSR and has granted consent for the nurse to access his information.

Releasing Information

When the RA receives confirmation of authorization from the CSR to access Doc-1 and Doc-2 of Bob's information, it does the following:

1. It sends Bob's system ID and indexes of Doc-1 and Doc-2 to the CSP. The following code (Figure 6.21 a) was used in the implementation to enable the CSR to search for patient's documents using the indexes.
2. It sends Bob's system ID and trapdoor tags to the SKA. The following code (Figure 6.21 b) was used in the implementation to enable the SKA to search for the trapdoors and send them to the user.

```

public class CloudServiceProvider {
    public CloudServiceProvider() {
    }
    public boolean sendDocsToClient(int numberOfDocs, String[] docIndexes, String[] outputDocNames) {
        try {
            S3Sample s3 = new S3Sample();
            File f=null;
            for(int i=0;i<numberOfDocs;i++)
            {
                f=s3.getDocument(docIndexes[i]);
                //Thread.sleep(500);
                s3.PutDocumentToOut(outputDocNames[i], f);
            }
            s3.listObjectInBuckt("docs-out");
            return true;
        } catch (Exception e) {
            // TODO: handle exception
            System.out.println("CloudServiceProvider.sendDocsToClient()");
            System.out.println(e.getMessage());
            return false;
        }
    }
}
a

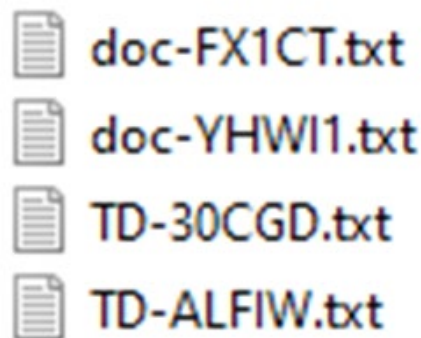
public class SecretKeyAgent {
    public SecretKeyAgent() {
    }
    public boolean sendKeysToClient(int numberOfKeys, String[] keyIndexes, String[] outputKeyNames) {
        try {
            S3Sample s3 = new S3Sample();
            File f=null;
            for(int i=0;i<numberOfKeys;i++)
            {
                f=s3.getTrapdoor(keyIndexes[i]);
                //Thread.sleep(500);
                s3.PutTrapdoorToOut(outputKeyNames[i],f );
            }
            s3.listObjectInBuckt("trapdoors-out");
            return true;
        } catch (Exception e) {
            // TODO: handle exception
            System.out.println("SecretKeyAgent.sendKeysToClient()" + e.getMessage());
            return false;
        }
    }
}
b

```

Figure (6.21) Codes to CSR and SKA

Decrypting Information

In response to the requests received from the RA, the CSP searches for Doc-1 and Doc-2 using their indexes and send them to the user. And the SKA does the same for the trapdoors and sends them to the user. The user then has two encrypted documents and two trapdoors. The user application associates trapdoors to their corresponding documents using the string exact match mechanism. And the secret keys in the trapdoors are then used to decrypt the documents.



(Figure 6.22 a) Indexes of information that is stored on the CSP

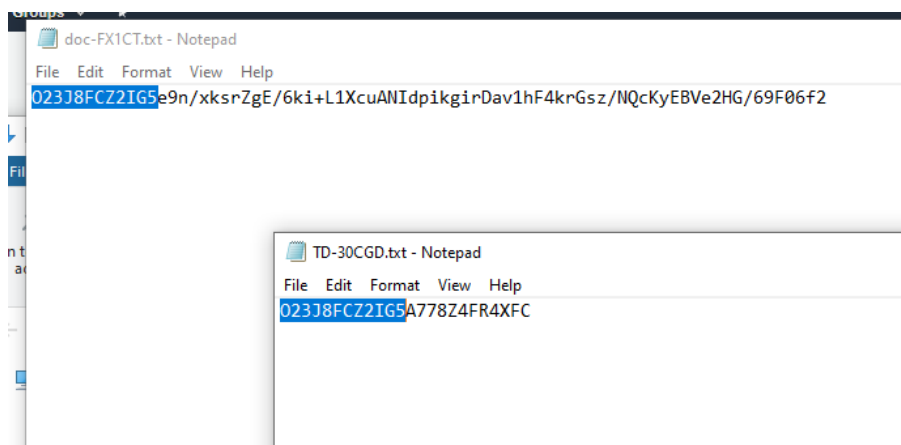


Figure (6.22 b) Text Exact Match

Testing the architecture

One of the main requirements of storing healthcare information on the cloud is the protection from unauthorized users. The cloud architecture was tested in terms of its ability to prevent unauthorized cloud users from accessing the information.

TEST 1: A request was made by a cloud user to download a trapdoor that is stored on the cloud.



Figure (6.23) Unauthorized attempt to down a trapdoor

RESULT

The access was denied, and the following message was shown

```
<Error>
  <Code>AccessDenied</Code>
  <Message>Access Denied</Message>
  <RequestId>DE119FB885660BEB</RequestId>
  <HostId>
    09mKJvt1AEkWenDpWJ97VwVPf2pluAwwTP82j86iyb58kKZDG/HDwQWM7mIq/Um9GMbHxHSS87Y=
  </HostId>
</Error>
```

Figure (6.24) Access denied in response to unauthorized action

TEST 2: Cloud users downloaded a document to read.

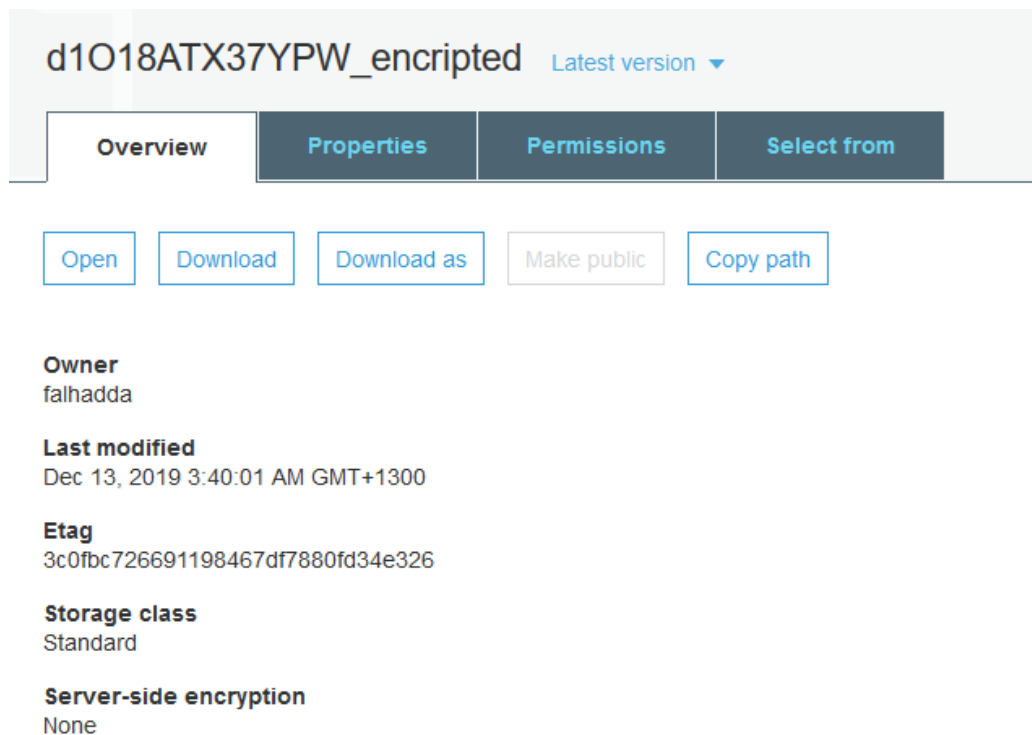


Figure (6.25) Downloading an encrypted document

RESULT

The document was viewed in its encrypted form as seen in Figure (2.26). The cloud user could not read its content. This means compromising the database of the CSP will always be fruitless to any disparate party, because the information is completely encrypted

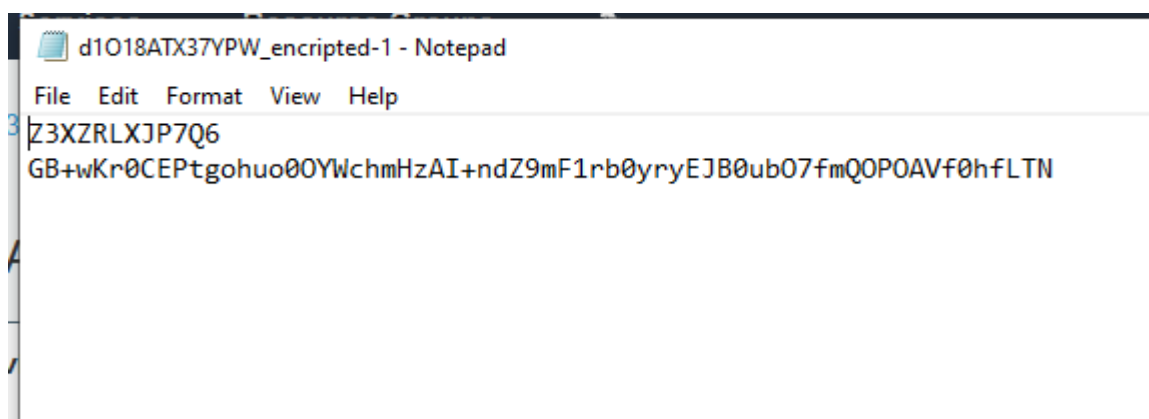


Figure (6.26) Unreadable encrypted document

TEST 3: Cloud users try to access the server or the database by sending queries

A cloud user was added as a normal cloud user and a query to access the system was made through its account.

RESULT

Access was denied

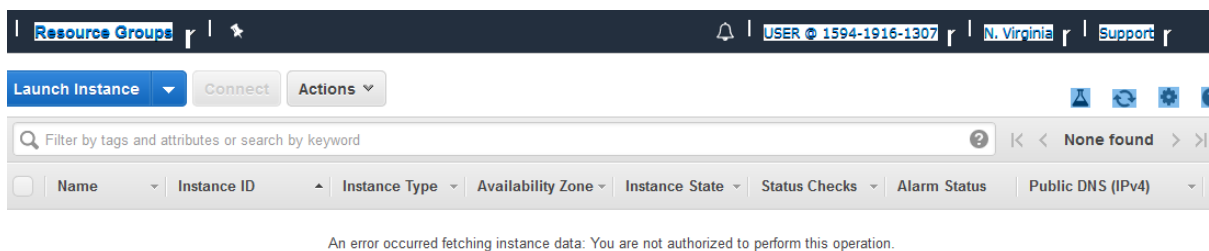


Figure (6.27) Response to unauthorized operation

TEST 4: Unregistered user tries to log in to the system

RESULT

Login failed as seen in Figure (6.28)

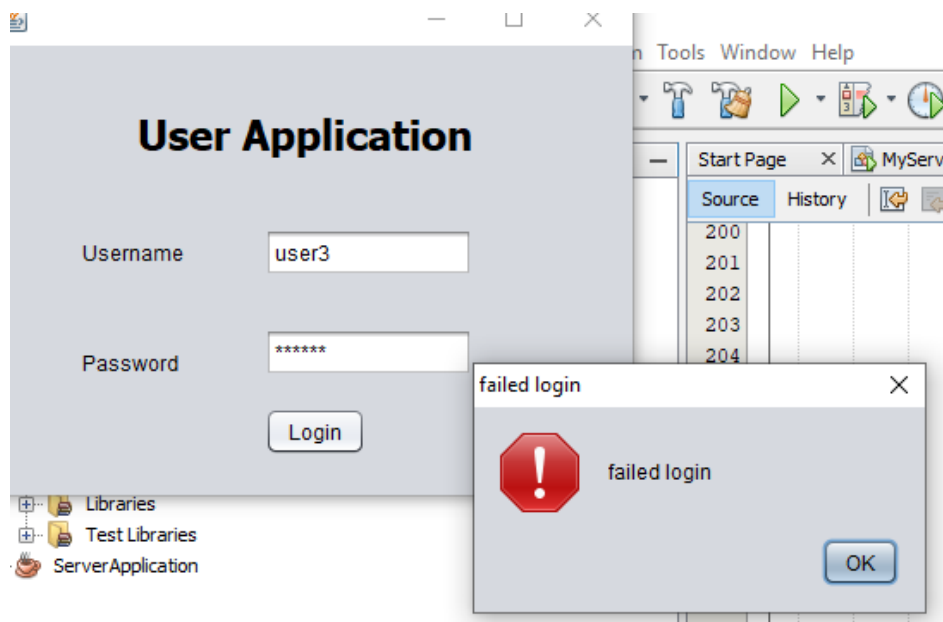


Figure (6.28) Response to an unauthenticated user

6.4 The use of patient records for research purposes

This section aims to present an instantiation of the proposed system's protocol for using patients' information for research purposes without violating the privacy of individual patients. The main objective of this section is to illustrate how the privacy-preserving strategies are incorporated into the proposed system design for sharing patients' records for research purposes. The first subsection illustrates how patients' records are accessed by disparate users (researchers), while the second subsection presents how the information is released for research purposes without violating the privacy of individual patients. Figure (6.29) illustrates the proposed system design and the interactions of its components.

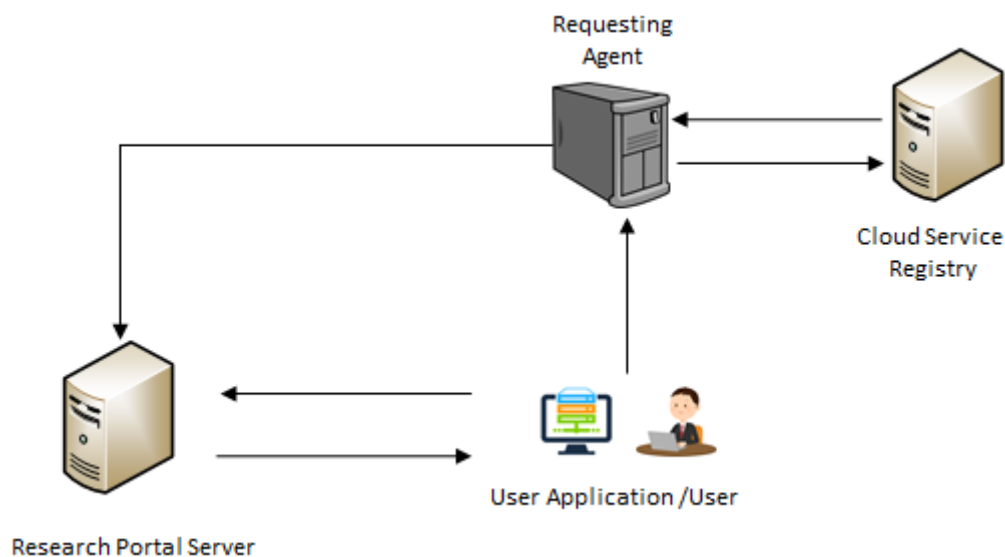


Figure (6.29) Sharing patients records for research purposes

For accessing the Research Portal Server (RPS), researchers are required to be authenticated and authorized. Authentication and authorization are granted by the CSR. Researchers send their requests to the RA. The RA authenticates users forward their requests to the CSR for authorization. Users are expected to have been enrolled in the system as “Researchers” prior to sending requests to the RA for accessing patient records. Similar to the SLAs granted to medical practitioners and users who have access to patients' records stored on the cloud; researchers are also granted access right to access the RPS in the form of SLAs.

The following section presents the protocol followed in the proposed system design for accessing the RPS and requesting datasets for research purposes. The illustration is presented

in a step-by-step fashion. The assumption made in the illustration that a researcher (Bob) is enrolled in the system as a researcher and granted access to the RPS.

6.4.1 Accessing the Research Portal Server

The proposed system employs a protocol that comprises three stages namely: Requesting Access, Authentication, and Authorization, and finally Initializing RSP – UA connection. The protocol aims to assure that accessing patients' information stored on the RSP is always authenticated and secure.

Stage 1: Requesting Access

Bob sends a request through the user application to the RA to access patients' records for research purposes. For this, Bob should be logged in the system as a researcher and clicks on the **RESEARCH** button available on the UA interface.

Stage 2: Authentication and Authorization

The RA authenticates Bob and forwards his request to the CSR for authorization. The CSR then confirms Bob's authorization to the RA to access the RPS.

Stage 3: Initializing RSP – UA Connection

When the RA receives the confirmation of the CSR, it sends a request to the RPS to activate a connection with Bob's UA. The request includes the CSR confirmation of authentication and authorization, as well as the physical address of Bob's machine. The RSP then activates a duplex connection with Bob's UA and confirms its availability to respond to Bob's queries.

Upon receiving confirmation from the RPS, the user application on Bob's machine is connected to the RPS and several functionalities and characteristics (described in section 5.2.1.2) are enabled on it. This makes the user application in a connected status and able to send queries and interact with the RPS.

Further, a list of attributes appears on the application interface. These attributes are for Bob to select from as the required attributes in the requested dataset. More information about this is presented in the following section.

6.4.2 Releasing patients' records

As soon as the UA is connected with the RPS, a list of attributes appears on the user interface, these attributes represent the attributes included in all patient information that is stored on the

RRS. Since patient information is stored in a standardized manner, the presentation of it is standardized. All patient information includes similar attributes that vary in their values. Therefore, these attributes appear on the UA interface for researchers to select from as the required attributes in the dataset needed for the intended research. However, there are a number of operations performed on patients' information (datasets) by the RPS before releasing them in response to the researcher's query. This section presents the process of requesting and obtaining patient datasets for research purposes in a step-by-step fashion.

6.4.3 Demonstration Setup

For this instantiation, a randomly generated dataset has been used. The dataset was generated using the built-in MS Excel's functionalities and contained 20,000 records, see (Appendix E). Every record is composed of twelve attributes (columns), and every attribute for every record was given a value that was randomly selected from a range of values assigned for that particular attribute. Table (6.1) presents a description of the dataset attributes used for this instantiation.

Attribute	Range of Values
Number	1 to 20000
Zip Code	A range of zip codes ranging between 155214 - 910041
Age	A range of ages between 1 to 100
Gender	Male / Female
Nationality	Random nationalities such as German, American Indian, American, African, Asian, Pacific, Black, White, Latino, ... etc
Marital Status	Married, Single, Engaged, Widowed, Separated
Blood Type	Random range of A+ , A- , B+ , B- , AB+ , AB- , O+ , O-
Medical Condition	A range of medical conditions such as Diabetes, Heart disease, Colon Cancer, Lung cancer, High or low Blood pressure, Blood cancer, none, ... etc.
Severity	Range from 1 – 5
Treatment	A range of dummy abbreviations such as CRG, ERF, BEU, XRF, TCD, GRT, DCU, CCR, VTT, CGD, UV, CWD, CER ... etc.
Donor/Non-donor	Donor, Non-donor
Allergy	A range of allergy types such as Insect sting allergy, cat allergy, Milk allergy, peanut allergy, Latex allergy, Eye allergy, ... etc.

Table (6.1) Description of the dataset used for this instantiation

The main intention of using a randomly generated dataset was to elaborate on how the proposed system design interactively releases patient records for research purposes in a privacy-preserving manner. Furthermore, the instantiation involved using MS SQL Server Management Studio (Microsoft, 2019) for demonstrating two operations involved in the process namely Dataset Generation and Value Transformation operations, while ARX software (Prasser, Kohlmayer, Lautenschläger, & Kuhn, 2014) was used to demonstrate the Value Generalization

operation and applying the (c, ℓ) -diversity model on the dataset prior to releasing it for research purposes.

SQL Server Management Studio (SSMS) is an open-source software application that is free-to-use. It is a management software that is used to connect with SQL Server and execute operations on SQL Server. The main goal of using SSMS was to demonstrate how the operations involved in the process of releasing patients' information for research purposes can be performed in the form of executable queries on a dataset. The dataset was imported to the SSMS to execute SQL queries on it using a localhost connection to the main server.

ARX is a data anonymization tool that effectively implements a range of privacy methods. It offers a programming interface for integration into other software systems and provides an intuitive cross-platform graphical interface. Moreover, ARX is a well-documented software that made it understandable to the researcher and accessible to apply privacy models on the dataset used. Therefore, ARX was used for elaborating the generalization operation on quasi-identifiers in the dataset and for applying the (c, ℓ) -diversity model on the dataset before releasing it for research purposes.

6.4.4 Implementation objectives

The main goal of the demonstration was to elaborate on how patient records are released in the proposed system design with elimination to the risk of probabilistic attacks that researchers may intentionally or unintentionally perform when looking at the released dataset. In this demonstration, the assumption made is that Bob is a researcher who requires accessing patients' information for research purposes. The main intention of Bob's research is to find out the influence of age on certain medical conditions namely diabetes, cancer, and High blood pressure. Therefore, Bob requires accessing records of patients who are diagnosed with diabetes, cancer, and high blood pressure.

Requesting patient information

Sending a query requires filling out an application form that appears on the UA interface as soon as it is connected to the RPS. The application has two parts: Attributes selection and value specifications.

Part 1: Attribute Selection

Bob is required to select the attributes that he needs to have in the required dataset. For this, he selects the following attributes which satisfy the need for his research:

Medical Conditions	Gender	Age	ZipCode	Treatment	Severity	Blood Type	Allergies
-----------------------	--------	-----	---------	-----------	----------	---------------	-----------

Part 2: Value Specifications

After the attributes have been selected, Bob is required to specify the target values of the medical condition attributes, as well as the level of generalization on the quasi-identifiers which are Zip code and Age.

The specification of requirements happens by selecting options from drop lists that appear when clicking on each of the selected attributes. The values of the “Medical Conditions” selected by Bob are Cancer, High blood pressure, and diabetes.

a) Age

Bob is required to specify the age intervals that are useful for his research. The minimum intervals allowed on the age attribute are: [5 years] for individuals who are 3 years old or above e.g. Age [3-8] years, and [3 months] for individuals who are below than 3 years old e.g. Age [12-15] months. However, Bob may require the age value to be generalized into larger intervals, which increases the anonymity of the dataset. In this instance, the assumption is that Bob requires the age to be presented in intervals of [10].

b) ZipCode

The zip code may have significant importance in some researches and less in others. The question about the importance of the “zip code” attribute value aims to find out the level of generalization allowed to the zip code value in the released dataset. The zip codes refer to geographical areas in which patients live, therefore, if a researcher is satisfied by grouping zip codes into large geographical areas such as an entire city or state, it becomes easier to generate more tuples that have a similar combination of quasi-identifiers leading to stronger anonymity

level of the released dataset. This question appears on the user application if the attribute “Zip Code” is selected as needed in the dataset required. In this instance, the assumption is that Bob requires the zip code to be in 2 characters which indicate to an entire city.

Summary of Bob’s Query

Deriving from the above, Bob needs to access patient information to conduct research that aims to find out the influence of age on a number of medical conditions which are diabetes, cancer, and High blood pressure. The attributes required are selected and the RPS will only release a dataset that contains the selected attributes according to the specifications provided by Bob. Below is a summary of Bob’s request that is sent to the RPS.

Medical Conditions	Gender	Age	ZipCode	Treatment	Level of Severity	Blood Type	Allergies
Cancer	Required	[10] Intervals	XX****	Value	Value	Value	Value
Diabetes							
High Blood Pressure							

Bob’s request that is sent to the Research Portal Server

Releasing dataset for research

Upon receiving Bob’s request, the RPS performs a number of operations to assure releasing a useful anonymized dataset that satisfies the need for it for Bob’s research. The below figure illustrates the operations that the RPS performs on patient records prior to releasing them for research purposes.

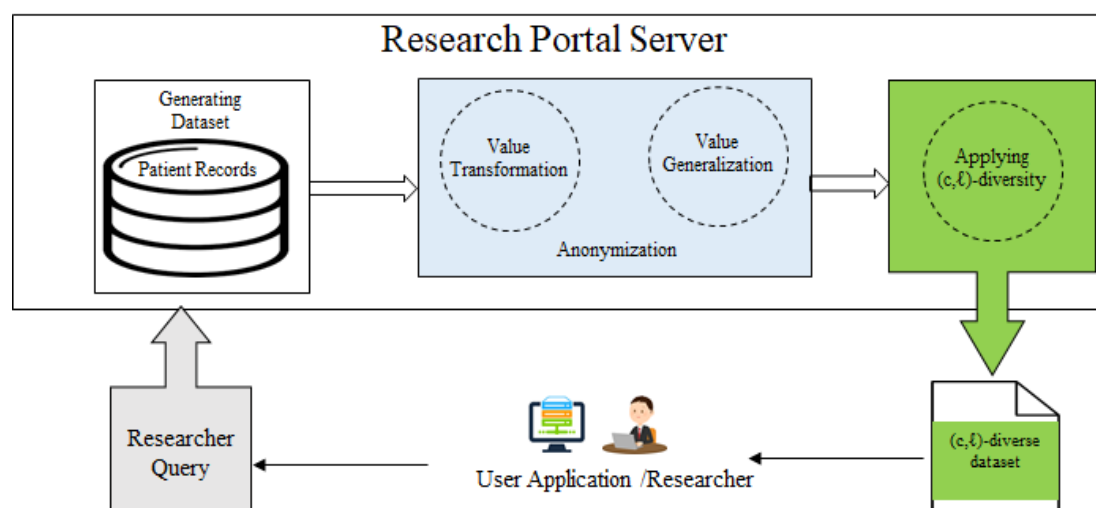


Figure (6.30) Releasing patients records for research purposes

As illustrated in figure (6.30), there are three operations performed by the RPS when responding to Bob's request for releasing dataset for research. These operations aim to produce an anonymized dataset that is useful for Bob's research.

1. Generating Dataset

Generating the dataset is the first operation performed by the RPS. It involves sampling tuples that comprise the required dataset. For this, the RPS employs an algorithm called Tuples Sampling Algorithm (TSA). The TSA algorithm aims to generate a dataset by selecting tuples from the stored patients' information according to two specific requirements, the first requirement is that the number of total tuples is no less than 3000, while the second requirement is that the number of records that have the target values for the target attributes is maximum 50% of the entire dataset. The target values in this instance are the values "Cancer", "Diabetes", and "High blood pressure" for the attribute "Medical Condition".

In response to Bob's query, the Tuples Sampling Algorithm employed by RPS generates a sample of tuples that contains at least 3000 tuples of which a maximum of 50% have the target values in the medical condition attributes. The input to the algorithm is the information that Bob has included in his query. Below is the input that the Tuples Sampling Algorithm takes in:

Target Attributes (T_a)	Medical Condition	Gender	Age	Zipcode	Treatment	Level of Severity	Blood Type	Allergies
Target Values (T_v)	Cancer	<i>Value (v)</i>	[10] Intervals	XX****	<i>Value (v)</i>	<i>Value (v)</i>	<i>Value (v)</i>	<i>Value (v)</i>
	Diabetes							
	Blood Pressure							

The TSA Algorithm generates the required dataset in two stages. Both stages involve retrieving tuples from the mother dataset (patient information) and copying them into a new table (dataset). The generated dataset undergoes anonymization operations before releasing it to Bob for conducting his research.

First Stage

The first stage involves retrieving tuples that meet Bob's requirements in terms of the required attributes and the target values. The retrieved tuples are copied into a new dataset which will be released to Bob after performing a number of anonymization operations on it. The required attributes in this instantiation are Medical Condition, Gender, Age, ZipCode, Treatment, Level of Severity, Blood Type, and Allergy, while the target values are specified in Bob's query namely 'cancer', 'diabetes', and 'blood pressure' for the Medical Condition attribute. Below is

a SQL example query that satisfies the above requirements for the first stage of generating the required dataset.

```

SELECT ZipCode, Age, Gender, BloodType, MedicalCondition,
Severity, Treatment, Allergy
INTO      dataset
FROM      Records
WHERE     MedicalCondition LIKE '%cancer%'
         OR MedicalCondition LIKE '%Diabetes%'
         OR MedicalCondition LIKE '%Blood Pressure%'

```

Zip code	Age	Gender	Blood type	Medical condition	Treatment	Severity	Allergy
155260	49	Female	A -	Cervical cancer	1	CT	Sulfa drugs allergy
155267	52	Female	O +	Oral cancer	2	VEF	none
155275	41	Male	A +	Lung cancer	1	CGD	Milk allergy
155279	17	Female	O +	Blood pressure	5	TUH	none
155287	20	Female	AB -	Cervical cancer	1	CCR	Dust allergy
155292	40	Female	AB +	Colon cancer	2	TCD	Dust allergy
155298	64	Male	AB -	Blood cancer	3	UUV	Eye allergy
155300	27	Male	O +	Blood pressure	1	ERF	Cockroach allergy
157803	94	Male	O +	Cervical cancer	2	XIN	Cockroach allergy
157806	67	Male	O +	Oral cancer	5	BRG	none
157808	39	Female	A +	Colon cancer	1	CCR	Latex allergy
157810	82	Female	AB +	Liver cancer	5	BIX	Insect sting allergy
157816	11	Male	B -	Liver cancer	3	CGD	none
157818	49	Female	AB -	Oral cancer	4	HUU	Rhinitis allergy
157819	88	Male	O +	Lung cancer	2	TTR	Insect sting allergy
157822	48	Male	O -	Diabetes	1	CFF	Insect sting allergy
157823	63	Male	O +	Diabetes	4	CFF	none
157825	64	Male	AB +	Blood cancer	3	GRT	Mold allergy
157826	17	Female	O +	Colon cancer	5	BTG	Aspirin allergy
157832	37	Male	O -	Colon cancer	1	CTV	Latex allergy
157835	41	Male	O +	Blood cancer	3	XCQ	Insect sting allergy
901774	48	Female	O -	Oral cancer	2	JED	Eye allergy
901779	84	Male	AB +	Oral cancer	4	BGG	Aspirin allergy
901790	83	Male	AB -	Oral cancer	4	HUU	Dust allergy
901805	65	Female	O -	Oral cancer	2	NHK	Peanut allergy
901806	35	Male	AB +	Lung cancer	3	XRF	Dust allergy
901811	67	Female	B -	Colon cancer	3	BIX	Cat allergy
901812	47	Male	O +	Colon cancer	2	GFF	Peanut allergy
158380	87	Male	B -	Liver cancer	5	CWD	Penicillin allergy
158381	69	Male	A -	Oral cancer	3	CWF	none
158385	96	Male	B +	Oral cancer	3	STT	none
158389	90	Female	A +	Oral cancer	1	CTA	Peanut allergy
158393	7	Male	A +	Colon cancer	4	NGF	Aspirin allergy
158395	1	Female	AB +	Blood pressure	1	CTX	Insect sting allergy
158401	55	Female	B +	Blood pressure	3	CTA	Chemotherapy drugs allergy
158404	94	Male	AB +	Oral cancer	2	UV	Latex allergy
158406	64	Male	B +	Diabetes	1	LID	Mold allergy
158408	53	Female	A +	Diabetes	3	VTT	Peanut allergy
158409	79	Female	B +	Liver cancer	3	DCU	Rhinitis allergy
158410	75	Male	AB -	Blood cancer	1	XIN	Skin allergy
158411	10	Male	A +	Blood pressure	3	VEX	none
158413	97	Male	A +	Blood pressure	2	UV	none
158417	95	Female	AB -	Oral cancer	1	TCD	Peanut allergy
902352	19	Female	A -	Oral cancer	4	CER	none
902353	100	Male	AB +	Liver cancer	1	NHJ	Rhinitis allergy
902356	31	Female	A +	Blood cancer	2	TTR	none
902360	35	Female	A -	Diabetes	1	BIX	none
902361	15	Female	AB +	Lung cancer	4	XEE	Rhinitis allergy

Table (6.2) Sample of the dataset tuples generated in the first stage

The outcome of this stage is a newly generated table called (dataset) which contains all tuples taken from the mother table (patient records) that have cancer, diabetes, or blood pressure as the values for the Medical Condition attribute. Moreover, all tuples in the dataset contain the eight attributes namely Zip Code, Age, Gender, Blood Type, Medical Condition, Severity, Treatment, and Allergy. Table (6.2) is a sample of tuples from the generated dataset in the first stage.

Second Stage

After generating a dataset that meets the requirements of Bob in terms of the required attributes and the target values, this stage involves inserting a number of tuples with values for the medical condition attribute that are different from the target values. The goal of this stage is to assure that at least 50% of the tuples in the dataset hold values that are different from the target values for the medical condition attribute. To achieve this, the Tuples Sampling Algorithm counts the tuples contained in the dataset generated in the first stage and accordingly decides the number of tuples to be inserted to it. The number of tuples in the generated dataset is referred to as the NumberOfTuples. Since it is a rule that no less than 3000 tuples to be released in any dataset for research purposes, the Tuples Sampling Algorithm decides the number of inserted tuples using the following rule:

```
IF      NumberOfTuples < 1500
THEN   TuplesInserted = 3000 - NumberOfTuples
Else   TuplesInserted = NumberOfTuples
```

Below is an example SQL query that satisfies the requirement of the second stage of generating the required dataset. The TotalTuples in the following query refers to the number of tuples in the mother dataset (patients' record) which is used in the query for technical purposes related to SQL Server.

INSERTINTO dataset

SELECTTOP **NumberOfTuples** ZipCode, Age, Gender, BloodType,
Medical Condition, Severity, Treatment, Allergy

FROM Records **TABLESAMPLE**(TotalTuples rows)

WHERE MedicalCondition **NOT LIKE** '%cancer%'

AND MedicalCondition **NOT LIKE** '%Diabetes%'

AND MedicalCondition **NOT LIKE** '%Blood Pressure%'

The insertion of tuples into the dataset results in having a dataset with at least 50% of tuples that have different values from the target ones.

ZipCode	Age	Gender	BloodType	MedicalCondition	Severity	Treatment	Allergy
155260	49	Female	A -	Cervical cancer	1	CT	Sulfa drugs allergy
155267	52	Female	O +	Oral cancer	2	VEF	none
155275	41	Male	A +	Lung cancer	1	CGD	Milk allergy
155279	8	Female	O +	Blood pressure	5	TUH	none
155287	20	Female	AB -	Cervical cancer	1	CCR	Dust allergy
155298	64	Male	AB -	Blood cancer	3	UUV	Eye allergy
155300	27	Male	O +	Blood pressure	1	ERF	Cockroach allergy
157803	94	Male	O +	Cervical cancer	2	XIN	Cockroach allergy
157819	88	Male	O +	Lung cancer	2	TTR	Insect sting allergy
157822	48	Male	O -	Diabetes	1	CFF	Insect sting allergy
157823	63	Male	O +	Diabetes	4	CFF	none
157825	64	Male	AB +	Blood cancer	3	GRT	Mold allergy
157826	9	Female	O +	Colon cancer	5	BTG	Aspirin allergy
157832	37	Male	O -	Colon cancer	1	CTV	Latex allergy
157835	41	Male	O +	Blood cancer	3	XCQ	Insect sting allergy
151774	48	Female	O -	Oral cancer	2	JED	Eye allergy
151779	84	Male	AB +	Oral cancer	4	BGG	Aspirin allergy
151790	83	Male	AB -	Oral cancer	4	HUU	Dust allergy
151805	65	Female	O -	Oral cancer	2	NHK	Peanut allergy
151806	35	Male	AB +	Lung cancer	3	XRF	Dust allergy
151811	67	Female	B -	Colon cancer	3	BIX	Cat allergy
151812	47	Male	O +	Colon cancer	2	GFF	Peanut allergy
158380	87	Male	B -	Liver cancer	5	CWD	Penicillin allergy
158381	69	Male	A -	Oral cancer	3	CWF	none
155214	89	Female	B -	Heat Stress	3	CRG	none
155217	98	Male	O +	Hepatitis B	3	XRF	Latex allergy
155218	88	Female	B +	Back Belts	5	TCD	Ibuprofen allergy
155219	31	Female	O -	Birth Defect	2	GRT	Cat allergy
155221	49	Male	B +	Hepatitis A	5	CCR	Cockroach allergy
155227	7	Female	A -	Yellow fever	4	CGD	Mold allergy
155228	54	Male	O +	Appendictis	3	SSW	Skin allergy
155230	9	Female	AB -	Brainerd Diarrhea	5	BTG	Dust allergy
155232	5	Female	A -	Hepatitis A	1	CQQ	Latex allergy
155233	5	Male	A -	Hapetitis C	4	GFF	Mold allergy
155234	13	Male	B +	Hearing impairment	2	STT	Skin allergy
155235	52	Male	B -	Appendictis	5	JED	none
155236	52	Female	B -	None	2	EFT	Aspirin allergy
155237	68	Male	O +	Hapetitis C	4	CT	Milk allergy
155240	10	Male	B -	Yellow fever	1	CWD	Rhinitis allergy
155241	31	Male	B -	Birth Defect	3	EBB	Milk allergy
155242	27	Female	O -	Back Belts	3	NHJ	Penicillin allergy
155244	39	Female	B +	Birth Defect	4	XRF	Sulfa drugs allergy
155246	14	Female	B +	Yellow fever	3	UUV	Mold allergy
155247	28	Male	B -	Yellow fever	4	XIN	Sulfa drugs allergy
155248	51	Female	A -	Hapetitis C	3	XCQ	Milk allergy
155250	15	Female	O +	Malaria	2	XIN	Insect sting allergy
155252	55	Male	O -	None	3	NHK	Ibuprofen allergy
155254	10	Female	A +	Back Belts	4	BGG	Chemotherapy drugs allergy

Table (6.3) Sample of the generated dataset after the insertion of tuples in stage 2

Table (6.3) presents a sample of tuples from the dataset generated after the insertion of tuples. As seen in table (6.3), there are 50% of tuples in the dataset contain values in the target attribute (Medical Condition) attribute that are different from the target values.

2. Anonymization

The anonymization stage aims to assure that patients' records in the generated dataset are anonymized before releasing the dataset to the researcher (Bob). The anonymization operation involves two operations that are performed on the dataset namely value transformation and value generalization. The transformation operation aims to completely change values of attributes, while the generalization operation aims to group values into more general ones.

a. Value Transformation

This operation aims to transform the non-targeted values to a different value that is not-useful-but-correct, such as "none", or "other", or "healthy". The transformation operation is performed on the values that are included in the tuples inserted in the second stage of the dataset generating process. Such transformation does not affect the utility of the dataset either its correctness, but the existence of tuples that have not-useful-but-correct values eliminates the possibility of successful speculations for direct disclosure. Therefore, 50% of the tuples in the dataset have the target values for the "Medical Condition" attribute and values of other medical condition-related attributes such as "Treatment" and "Severity", while the rest of tuples have a non-useful-but-correct value such as (none) for the same attributes. Below is an example query that is run by the RPS for the value transformation operation.

```
UPDATE dataset SET MedicalCondition ='none ',  
                Severity  ='',  
                Treatment ='none'  
WHERE MedicalCondition NOT LIKE '%cancer%'  
      AND MedicalCondition NOT LIKE '%Diabetes%'  
      AND MedicalCondition NOT LIKE '%Blood Pressure%'
```

The query aims to replace the values of the attributes ‘Medical Condition’ and ‘Treatment’ by “none”, and blank for the Severity attribute for all tuples that do not have any of the target values. The value “none” refers to other medical conditions that are not included in the range of the target values (cancer, high blood pressure, and diabetes). It could also mean that individuals who have the value “none” for the medical condition do not suffer from any disease. Therefore, the existence of tuples that have the value ‘none’ eliminates the possibility of any attempt for direct disclosure that Bob may make when looking at the dataset.

Table (11) presents the sample of the generated dataset tuples after performing the value transformation operation on the dataset.

ZipCode	Age	Gender	BloodType	MedicalCondition	Severity	Treatment	Allergy
155260	49	Female	A -	Cervical cancer	1	CT	Sulfa drugs allergy
155267	52	Female	O +	Oral cancer	2	VEF	none
155275	41	Male	A +	Lung cancer	1	CGD	Milk allergy
155279	8	Female	O +	Blood pressure	5	TUH	none
155287	20	Female	AB -	Cervical cancer	1	CCR	Dust allergy
155298	64	Male	AB -	Blood cancer	3	UUV	Eye allergy
155300	27	Male	O +	Blood pressure	1	ERF	Cockroach allergy
157803	94	Male	O +	Cervical cancer	2	XIN	Cockroach allergy
157819	88	Male	O +	Lung cancer	2	TTR	Insect sting allergy
157822	48	Male	O -	Diabetes	1	CFF	Insect sting allergy
157823	63	Male	O +	Diabetes	4	CFF	none
157825	64	Male	AB +	Blood cancer	3	GRT	Mold allergy
157826	9	Female	O +	Colon cancer	5	BTG	Aspirin allergy
157832	37	Male	O -	Colon cancer	1	CTV	Latex allergy
157835	41	Male	O +	Blood cancer	3	XCQ	Insect sting allergy
151774	48	Female	O -	Oral cancer	2	JED	Eye allergy
151779	84	Male	AB +	Oral cancer	4	BGG	Aspirin allergy
151790	83	Male	AB -	Oral cancer	4	HUU	Dust allergy
151805	65	Female	O -	Oral cancer	2	NHK	Peanut allergy
151806	35	Male	AB +	Lung cancer	3	XRF	Dust allergy
151811	67	Female	B -	Colon cancer	3	BIX	Cat allergy
151812	47	Male	O +	Colon cancer	2	GFF	Peanut allergy
158380	87	Male	B -	Liver cancer	5	CWD	Penicillin allergy
158381	69	Male	A -	Oral cancer	3	CWF	none
155214	89	Female	B -	none	none	none	none
155217	98	Male	O +	none	none	none	Latex allergy
155218	88	Female	B +	none	none	none	Ibuprofen allergy
155219	31	Female	O -	none	none	none	Cat allergy
155221	49	Male	B +	none	none	none	Cockroach allergy
155227	7	Female	A -	none	none	none	Mold allergy
155228	54	Male	O +	none	none	none	Skin allergy
155230	9	Female	AB -	none	none	none	Dust allergy
155232	5	Female	A -	none	none	none	Latex allergy
155233	5	Male	A -	none	none	none	Mold allergy
155234	13	Male	B +	none	none	none	Skin allergy
155235	52	Male	B -	none	none	none	none
155236	52	Female	B -	none	none	none	Aspirin allergy
155237	68	Male	O +	none	none	none	Milk allergy
155240	10	Male	B -	none	none	none	Rhinitis allergy
155241	31	Male	B -	none	none	none	Milk allergy
155242	27	Female	O -	none	none	none	Penicillin allergy
155244	39	Female	B +	none	none	none	Sulfa drugs allergy
155246	14	Female	B +	none	none	none	Mold allergy
155247	28	Male	B -	none	none	none	Sulfa drugs allergy
155248	51	Female	A -	none	none	none	Milk allergy
155250	15	Female	O +	none	none	none	Insect sting allergy
155252	55	Male	O -	none	none	none	Ibuprofen allergy
155254	10	Female	A +	none	none	none	Chemotherapy drugs allergy

Table (6.4) Sample of value-transformed tuples

As seen in the table (6.4), the dataset -after performing the value transformation operation on it- contains 50% of tuples that do not have transformed values of the Medical Condition,

Severity, and Treatment attributes. The inclusion of such tuples makes Bob require more pieces of information to be able to identify an individual or associate a medical condition to an individual. In this scenario, if Bob wants to identify any individual, the percentage of his successful speculation is low, and increasing it requires knowing the blood type of the individual. Nevertheless, having a dataset that does not contain information about all patients enrolled in the system is another privacy-preserving characteristic of the released dataset, because for identifying an individual patient or associating a particular tuple with an individual patient, it is first required to know whether the patient is included in the released dataset or not, therefore, it is difficult for Bob to confidently reach any conclusion related to patient identity or attribute.

b. Generalization

The generalization operations are performed on quasi-identifiers for the goal of having several records that have similar combinations of quasi-identifiers in the dataset. The generalization operation in this scenario is performed on the age and zip code as per Bob's query. The generalization of the age refers to the process of replacing the age value with a less specific but semantically consistent value such as age intervals. For example, the ages 22, 23, 29, 20, and 30 can be replaced by an interval of {2030}. The zip code is generalized by the masking portion of its characters. For example, the zip code in this scenario is required to indicate to an entire city, therefore, only the characters that indicate to the entire city appear in the value of the zip code, for example, the zip code 155244 becomes 15**** after performing the generalization operation on it. The generalization operation in this demonstration was performed on the dataset using the ARX tool together with applying the (c, ℓ) -diversity that is explained in the following section. The age was generalized into 10 years age intervals, while the zip code was generalized by masking the first 4 characters and leave only 2 characters that indicate an entire city. Table (12) presents a sample of the dataset after performing the generalization operation on the dataset and applying the (c, ℓ) -diversity model.

3. Applying (c, ℓ) -diversity

Finally, the last operation performed on the dataset by the RPS before releasing it is applying the (c, ℓ) -diversity model. This operation aims to set a certain frequency of values occurrences in every q^* -block in the dataset. The main goal of this operation is to release a dataset in which at most 50% of tuples in every q^* -block hold the target values of the target attributes, while the rest of tuples have “non-useful-but-correct” values of the same. Therefore, to satisfy Bob's

query and release a useful dataset for his research, every q^* -block of the released dataset should contain at least 50% of tuples that have the value “none” in the Medical Condition attribute and other medical condition-related attributes. For that, the dataset released to Bob should be (2, 2)-diverse. Table (6.5) represents (2,2)-diverse dataset that is released in response to Bob’s query. The parameter c refers to the frequency of occurrence for the target values in the “Medical condition” attribute and its related attributes.

ZipCode	Age	Gender	Blood type	Medical condition	Treatment	Severity	Allergy
48****	{10 20}	Female	B -	Blood pressure	3	XIN	none
48****	{10 20}	Female	B -	Cervical cancer	3	KLJ	Dust allergy
48****	{10 20}	Female	B +	Lung cancer	3	CT	Skin allergy
48****	{10 20}	Female	O -	none	0	none	none
48****	{10 20}	Female	B -	none	0	none	none
48****	{10 20}	Female	AB -	none	0	none	Sulfa drugs allergy
48****	{20 30}	Female	O +	Colon cancer	2	XEE	Milk allergy
48****	{20 30}	Female	B -	none	0	none	Sulfa drugs allergy
48****	{40 50}	Female	O +	Blood cancer	5	CGD	Peanut allergy
48****	{40 50}	Female	O -	Cervical cancer	2	CT	none
48****	{40 50}	Female	A -	Diabetes	2	TCS	Peanut allergy
48****	{40 50}	Female	B -	Liver cancer	3	CRG	none
48****	{40 50}	Female	B +	none	0	none	none
48****	{40 50}	Female	O +	none	0	none	none
48****	{40 50}	Female	AB -	none	0	none	none
48****	{60 70}	Female	B +	Diabetes	4	PIN	Rhinitis allergy
48****	{60 70}	Female	AB -	none	0	none	none
48****	{60 70}	Female	AB -	none	0	none	Insect sting allergy
48****	{60 70}	Female	AB +	none	0	none	Insect sting allergy
48****	{10 20}	Male	AB -	Diabetes	3	TUH	Sulfa drugs allergy
48****	{10 20}	Male	O +	Lung cancer	4	CRG	none
48****	{10 20}	Male	O -	none	0	none	none
48****	{10 20}	Male	B -	none	0	none	Skin allergy
48****	{40 50}	Male	A +	Blood cancer	1	CCT	Latex allergy
48****	{40 50}	Male	AB +	Diabetes	4	NHK	Ibuprofen allergy
48****	{40 50}	Male	A +	Lung cancer	2	ACT	Chemotherapy drugs allergy
48****	{40 50}	Male	B +	none	0	none	Ibuprofen allergy
48****	{40 50}	Male	A -	none	0	none	none
48****	{40 50}	Male	AB -	none	0	none	Milk allergy
48****	{40 50}	Male	A -	none	0	none	none
48****	{40 50}	Male	O +	none	0	none	Rhinitis allergy
48****	{50 60}	Male	AB +	Colon cancer	3	CRG	none
48****	{50 60}	Male	A +	Liver cancer	1	JED	none
48****	{50 60}	Male	A -	none	0	none	Sulfa drugs allergy
48****	{50 60}	Male	AB -	none	0	none	none
48****	{50 60}	Male	O -	none	0	none	Skin allergy
48****	{10 20}	Female	B -	Blood pressure	3	XIN	none
48****	{10 20}	Female	B -	Cervical cancer	3	KLJ	Dust allergy
48****	{10 20}	Female	B -	none	0	none	none
48****	{10 20}	Female	AB -	none	0	none	Sulfa drugs allergy
48****	{20 30}	Female	O +	Colon cancer	2	XEE	Milk allergy
48****	{20 30}	Female	B -	none	0	none	Sulfa drugs allergy
48****	{40 50}	Female	O +	Blood cancer	5	CGD	Peanut allergy
48****	{40 50}	Female	O -	Cervical cancer	2	CT	none
48****	{40 50}	Female	A -	Diabetes	2	TCS	Peanut allergy
48****	{40 50}	Female	B -	Liver cancer	3	CRG	none
48****	{40 50}	Female	B +	none	0	none	none
48****	{40 50}	Female	O +	none	0	none	none
48****	{40 50}	Female	AB -	none	0	none	none
48****	{40 50}	Female	A -	none	0	none	Ibuprofen allergy
48****	{60 70}	Female	AB +	Blood cancer	4	TCS	none
48****	{60 70}	Female	B +	Diabetes	4	PIN	Rhinitis allergy
48****	{60 70}	Female	AB -	none	0	none	none
48****	{60 70}	Female	AB -	none	0	none	Insect sting allergy
48****	{60 70}	Female	AB +	none	0	none	Insect sting allergy

Table (6.5) (2,2) - diverse dataset

Therefore, setting it to the value 2 means that in every q^* -block, the number of tuples that have the values “cancer”, or “high blood pressure” or “diabetes” appear at most $\frac{1}{2}$ of the total records, while the other half of the records hold the value “none” for the same attributes. Setting the parameter ℓ to 2 means that in every q^* -block, there should be at least 2 well-represented values of the medical condition attributes which are in Bob’s case the values that indicate to unhealthy medical conditions which are (“cancer”, “high blood pressure”, “diabetes”), and the value “none”.

As seen in Table (6.5), tuples in every q^* -block contains two types of values (unhealthy and none) for the medical condition attribute of which at least 50% of them are “none”. Such frequency of value occurrence eliminates the chance of successful speculations that Bob may make to directly associate an attribute to an individual. For example, using the below q^* -blocks which were taken from Table (6.5), if Bob knows that Alice lives in zip code 481142 and she is 43 years old, he can conclude with maximum 50% confidence that Alice has an unhealthy medical condition because 50% of the tuples hold the value none in the medical condition attribute and other related attributes.

48****	{40 50}	Female	O +	Blood cancer	5	CGD	Peanut allergy
48****	{40 50}	Female	O -	Cervical cancer	2	CT	none
48****	{40 50}	Female	A -	Diabetes	2	TCS	Peanut allergy
48****	{40 50}	Female	B -	Liver cancer	3	CRG	none
48****	{40 50}	Female	B +	none	0	none	none
48****	{40 50}	Female	O +	none	0	none	none
48****	{40 50}	Female	AB -	none	0	none	none
48****	{40 50}	Female	A -	none	0	none	Ibuprofen allergy
48****	{60 70}	Female	AB +	Blood cancer	4	TCS	none
48****	{60 70}	Female	B +	Diabetes	4	PIN	Rhinitis allergy
48****	{60 70}	Female	AB -	none	0	none	none
48****	{60 70}	Female	AB -	none	0	none	Insect sting allergy
48****	{60 70}	Female	AB +	none	0	none	Insect sting allergy

Moreover, the possibility of Alice being a Liver cancer patient is calculated by dividing the number of tuples that contain Liver cancer for the medical condition attribute, on the total tuples that have similar combinations of quasi-identifiers in the same q^* -block. Therefore, the chance of successful speculation that Bob may make about Alice -if he knows her record exists in the dataset- is maximum $\frac{1}{8}$.

The confidence of any speculations that Bob may make is in roughly inverse proportion to the number of tuples in the dataset that have similar combinations of quasi-identifiers. Therefore, Bob requires more helpful pieces of information to increase the chance of concluding the

medical condition of Alice. In this example, for Bob to be sure that Alice suffers from Liver cancer, he needs to know that her record exists in the dataset, and her blood type is (B-).

6.5 Summary

This chapter aimed to demonstrate how the proposed cloud architecture can satisfy the need for information sharing in the healthcare domain without questioning the privacy of patients' information. The chapter presented a demonstration of how the proposed architecture enables for storing and sharing healthcare information for both; medical treatment and research purposes without questioning the privacy of patients. For sharing healthcare information for medical treatment purposes, the cloud architecture was implemented using Amazon Web Services and a scenario of accessing stored patients' information was demonstrated. The ability of the cloud to store and share patients' information in a privacy-preserving manner was proven. The cloud architecture was primarily tested against many possible actions that could be performed on information that is stored on the cloud to breach the privacy of it.

Sharing healthcare information for research purposes is another important aspect of the proposed architecture. The strategies and techniques of the proposed architecture for preserving the privacy of patients' information when it is used for research purposes were tested. The proposed architecture performs several operations patients' information before releasing them for research purposes. These operations have been tested in terms of their efficiency in preserving the anonymity of the information without affecting the utility of it for the intended research purpose.

In conclusion, the instantiation and implementation of the proposed architecture have validated the proposed cloud architecture in terms of its concept and application in real cloud-based healthcare applications.

Chapter 7: Discussion

The purpose of this research was to design a cloud-based architecture for storing and sharing healthcare information in a privacy-preserving manner. The Design Science Research Methodology (DSRM) was employed for the conduction of this research. As illustrated in Figure (7.1), the activity of designing and developing the architecture was iterative with the evaluation activity. The evaluation of the architecture design was performed against the objectives of the solution defined in the second activity of the DSRM. The reason behind the iterations between activities was to guarantee that the architecture design met the objectives of the solution defined in this research.

The objectives of the solution were defined as results of case study data analysis and data

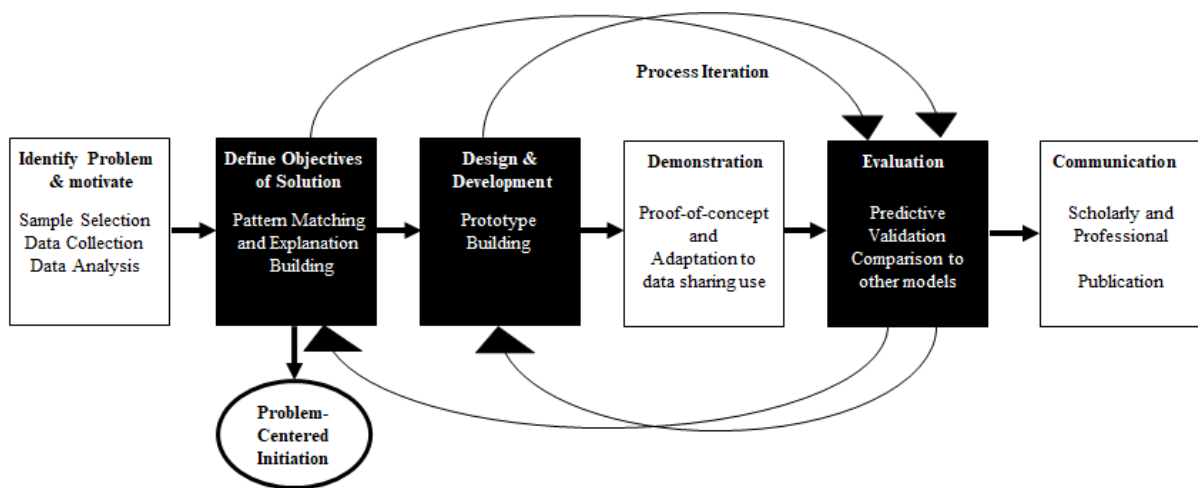


Figure (7.1) Activity Iterations

obtained from the literature. The purpose of the case study research was to inquire as to the way healthcare practitioners use shared information for medical care purposes. Data was collected through literature review and open-ended interviews with research participants. The explanation-building technique was employed to match the patterns extracted from the analysis of the data collected. The results of the pattern-matching indicated the characteristics of the intended architecture (objectives of solution).

In the previous chapter, the designed architecture was attested through scenario-based instantiations. The instantiations demonstrated how the designed architecture met the defined objectives of the solution for storing patients' information on the cloud and sharing it among genuine parties in a privacy-preserving manner. Therefore, the researcher has come to the point

that the designed architecture meets the objectives of the solution defined in the second activity of the DSRM.

The purpose of this chapter is twofold: firstly, it aims to evaluate the contribution of the research methodology employed for this research (DRSM) for thesis critical reflection. The goal is to explain the advantages of the methodology lifecycle in terms of overcoming research problems and achieving innovative outcomes. Secondly, for completeness of the research journey, it is important to identify the rationale between the findings of the research and the research questions using a Quasi-Judicial scholarly method. Figure (7.2) illustrates the flow of the research in light of the DSRM activities and their iterations.

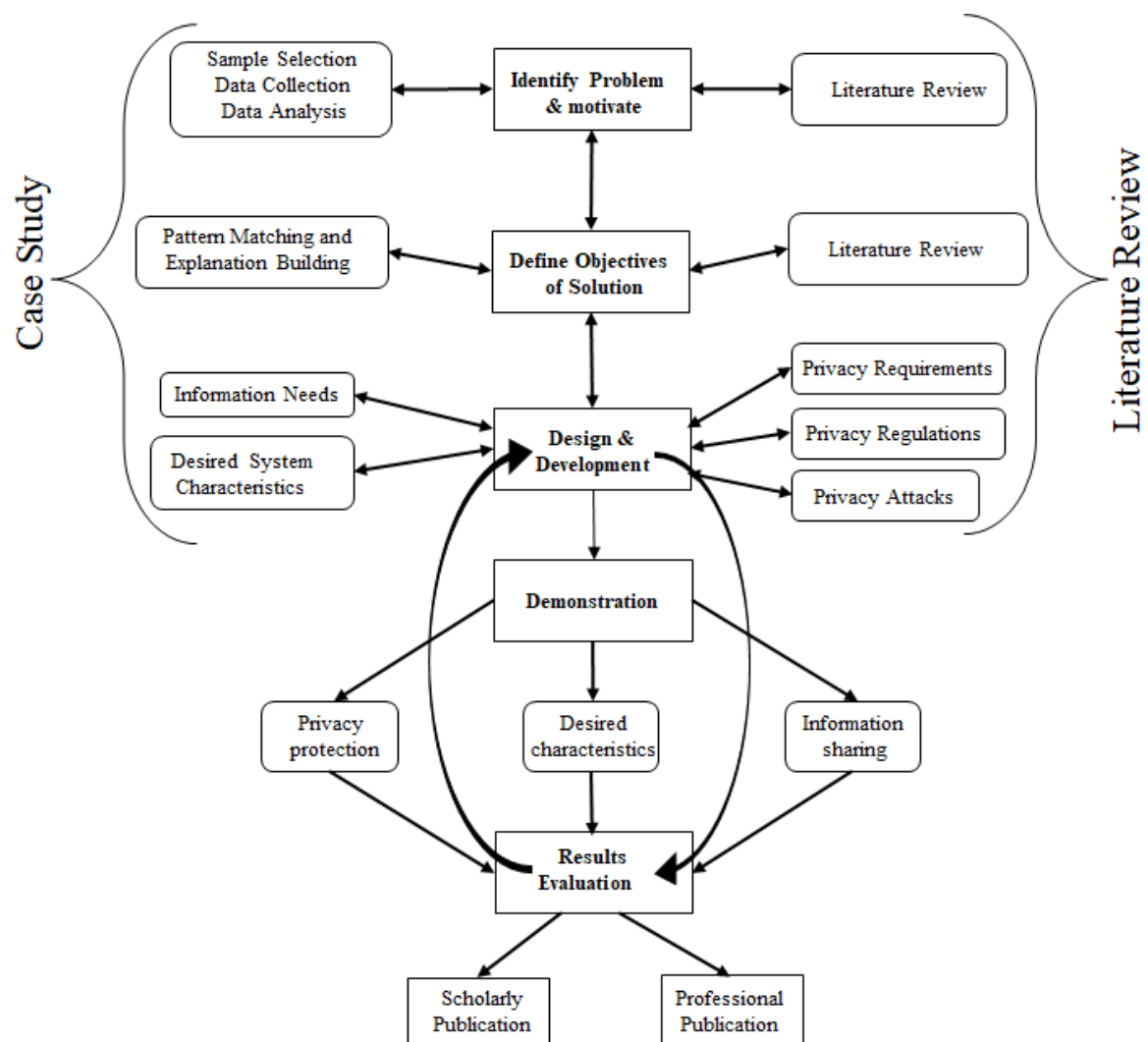


Figure (7.2) Research Flow in light of DSRM

This chapter is structured to take evidence from previous chapters and use them for qualitative testing. The chapter is organized as the following: Section 7.1 presents the research questions

were addressed, section 7.2 presents the implication of the results, section 7.3 presents the contribution of this research to the body of knowledge, while section 7.4 concludes the chapter.

7.1 Artefact Evaluation

The required characteristics of healthcare information systems have been identified in chapters four and five. In Chapter 4, the required user-related characteristics have been identified from the findings of the case study data analysis. Chapter 5 presented the privacy-related requirements -identified in the literature- for dealing with healthcare information. The researcher has reached to the point that the designed cloud architecture satisfies the need for storing and sharing healthcare information in a privacy-preserving manner. It is also believed that the desired characteristics identified in Chapter 4 are also met in the designed architecture. This section presents the evaluation of the designed cloud architecture in terms of its ability to share healthcare information in a privacy-preserving manner, as well as meeting the desired characteristics that have been identified in Chapter 4.

The main intention of designing the artefact was to enable the collaborative use of healthcare information in a privacy-preserving manner. The challenge was related to preserving the privacy of information while it is stored and shared in a cloud-based environment. The proposed cloud architecture will serve the healthcare domain by storing all patients' health information in one place (cloud) so that genuine users can access it regardless of their locations. The mechanisms followed in the designed architecture to store and share information were demonstrated in Chapter 6 in scenario-based instantiations. In the instantiation, it was confirmed that the designed architecture overcomes the challenges related to the privacy and confidentiality of information when it is stored and shared in cloud-based environments.

This section presents a discussion of how the designed artefact meets the objectives of the solution in light of the Design Science Research Methodology. The designed cloud architecture is discussed in terms of its ability to share information in a privacy-preserving manner, as well as meeting the desired characteristics.

7.1.1 Sharing Information

In Chapter 6, the ability of the designed architecture to facilitate collaborative use of healthcare information was elaborated in a scenario-based instantiation. The instantiation was presented in two parts: the first part illustrated the ability of the designed architecture to share healthcare information for medical treatment purposes, while the second part elaborated on how patients'

health-related information is stored and used for research purposes. Both parts aimed to illustrate how the designed cloud-based architecture enables for effective use of patients' information without breaching the privacy of patients or their information.

Sharing information for medical treatment purposes

The elaboration of the proposed architecture included three scenarios: (1) enrolling a patient in the system and storing his information, (2) requesting the patients' information for medical treatment purpose, and finally (3) updating his information. The proposed architecture was implemented in a real cloud context scenario. The ability of the proposed architecture to share healthcare information was tested. The validity of separating encrypted patients' information from their trapdoors was conceptually and technically proven. Patients' information could be accessed by a nurse according to pre-determined access rights. The following table presents a set of characteristics that were tested in the proposed architecture.

1	Patient' information can be stored on the cloud and collaboratively used in a privacy-preserving manner	Achieved
2	Information is disclosed according to access rights and nature of medical treatment for which patient information is required.	Achieved
3	Patients' have control over who can access their information	Achieved

Sharing information for research purposes

System instantiation of how the designed architecture works in terms of sharing healthcare information was presented in a scenario-based fashion. The proposed architecture was demonstrated in terms of its interactivity with researchers' queries. The main concern when using patients' information for research purposes is the privacy of the information. The proposed architecture was technically tested in terms of its ability to anonymize information without affecting the utility of it. The ability of the architecture to eliminate the privacy-related attacks on aggregated patients' information was proven. The following table presents the characteristics of the proposed architecture in terms of sharing healthcare information for research purposes.

1	Patient' information can be aggregated for research purposes	Achieved
2	Patients information is protected against Linkage Attacks	Achieved
3	Chances of successful probabilistic attacks on patient's information when released for research purposes are eliminated	Achieved

7.1.2 Privacy-preservation

One of the main issues of adopting cloud technology in the healthcare domain is the privacy and confidentiality of the information stored on the cloud. Storing sensitive information in the hand of a third party such as cloud providers has always been a major concern to the owners of information. Such concern has stemmed from the ability of cloud providers to read the information stored on the cloud without consent from the concerned parties mainly information owners.

In the proposed architecture, information is encrypted before it is sent to the cloud for storage. The cloud receives encrypted information to store without helpful information for decrypting it (secret keys). No decryption processes are performed on information while it is on the cloud. The searchable symmetric encryption scheme (SSE) employed in the proposed design assures that cloud providers can satisfy users' requests by releasing the required information without having to decrypt it, meaning that information is only decrypted while it is in the hand of genuine parties. The inability of cloud providers to read the information stored on the cloud overcomes the main privacy challenge of adopting cloud technology for storing healthcare information.

Information disclosure is another concern that hinders the adoption of cloud computing in the healthcare domain. Information that is stored in one place (cloud) is prone to complete disclosure by any disparate party. Patients may not accept their sensitive information to be disclosed for a medical practitioner who does not need such information in particular incidents, therefore, granting full access to information may breach the privacy of patients.

Dividing patients' information into many divisions according to the need of it overcomes the issue of information disclosure. The designed architecture categorizes patients' information into four categories of which three are for medical treatment purposes namely All_V, Em_V,

and OutP_V, and one is for research purposes (R). Medical practitioners do not always require accessing the entire information about a patient every time medical treatment is needed. For example, information about a sexual disease may not be needed in urgent medical treatments such as car accidents or minor incidents such as skin wounds and cuts. This was derived from the findings of the case study data analysis and supported by the literature. Therefore, dividing patients' information according to the need for it in different contexts is a solution to protect the privacy of information.

Nevertheless, the designed architecture requires obtaining patients' consent whenever accessing their information is required. The Cloud Service Registry (CSP) in the proposed architecture does not authorize users to access patients' information without having patients' consent. This gives patients means of control over who can access their health information which was identified as a requirement that patients wish to have. The following table presents the main privacy-preserving characteristics that were achieved in the proposed architecture.

1	Cloud provider cannot read patients' information stored in the database	Achieved
2	Information is only read by authorized users	Achieved
3	Just-enough Information is disclosed to authorized users in medical treatment incidents	Achieved
4	The proposed architecture adheres to HIPPA in terms of information privacy	Achieved

7.2 Research methodology contribution

The contribution of the mixed research methodology (Case Study and Design Science) is evaluated for this research. This section aims to elaborate on how the Design Science Research Methodology has contributed towards achieving the outcome of this research. The six activities, Problem Identification and motivation, Defining the Objectives of the Solution, Designing and development, Demonstration, Evaluation, and communication in this research are discussed.

Problem identification and motivation was the first activity performed in the design science research methodology followed in this research. This activity aimed to investigate the current

practices followed by healthcare information systems for sharing information. The goal was to understand how information is needed and how the current information systems employed by the healthcare domain can be improved. In this activity, the research inquired as to the way healthcare practitioners used shared information for medical care purposes. A sample of four medical institutions was selected for this research and a number of medical practitioners from each institution were invited to participate in the research. The researcher collected data for the research through open-ended interviews with the research participants and relevant literature. The collected data was then systematically filed and analyzed within the context of each individual case. Patterns were extracted from the data and the need for sharing healthcare information was understood in depth. The result of this activity was identifying the problem of the current healthcare information systems and realizing what was needed to overcome it.

The second activity involved **defining the objectives of the solution**. The main objective of this research was to propose a cloud-based architectural design for sharing healthcare information in a privacy-preserving manner. This activity involved employing an explanation-building technique to match the patterns extracted in the previous activity to the theoretical proposition of the study. The information obtained through the explanation-building technique together with the information obtained from the relevant literature has indicated the objectives of the solution. The objectives of solutions refer to characteristics that healthcare information systems should have to best serve the healthcare domain in terms of storing and sharing information in a privacy-preserving manner. Many system characteristics were identified in this activity. In Chapter three and Chapter four, a set of artefact's required characteristics were derived from the case study data analysis. These characteristics were related to how healthcare information systems could best serve the healthcare domain in terms of storing and sharing information from users' points of view. In Chapter five, another set of characteristics was derived from the relevant literature as the requirements for the successful adoption of cloud computing technology in the healthcare domain in terms of information privacy and confidentiality. Each characteristic was meant to be a solution to satisfy a need in terms of sharing healthcare information in a privacy-preserving manner. The characteristics identified in this research activity were considered in the design and development activity.

The third activity involved is **designing and developing the artefact**. The activity of designing the artefact (cloud-architecture) was performed based on the inputs from the previous activity. The characteristics identified in the previous activity were considered in the design of the artefact. The design included two levels of granularity; the privacy-related characteristics of

the architecture as identified in the relevant literature, and the satisfaction of the needs for storing and sharing information as identified in the explanation building technique in the previous activity. The search for an effective artefact requires utilizing available means to reach desired ends while satisfying laws in the problem environment (Hevner, March, Park, & Ram, 2004). The researcher in this activity attempted to draw from the current cloud-compatible privacy protection techniques proposed in the literature concerning the need for sharing healthcare information as identified in the explanation-building technique in the previous activity.

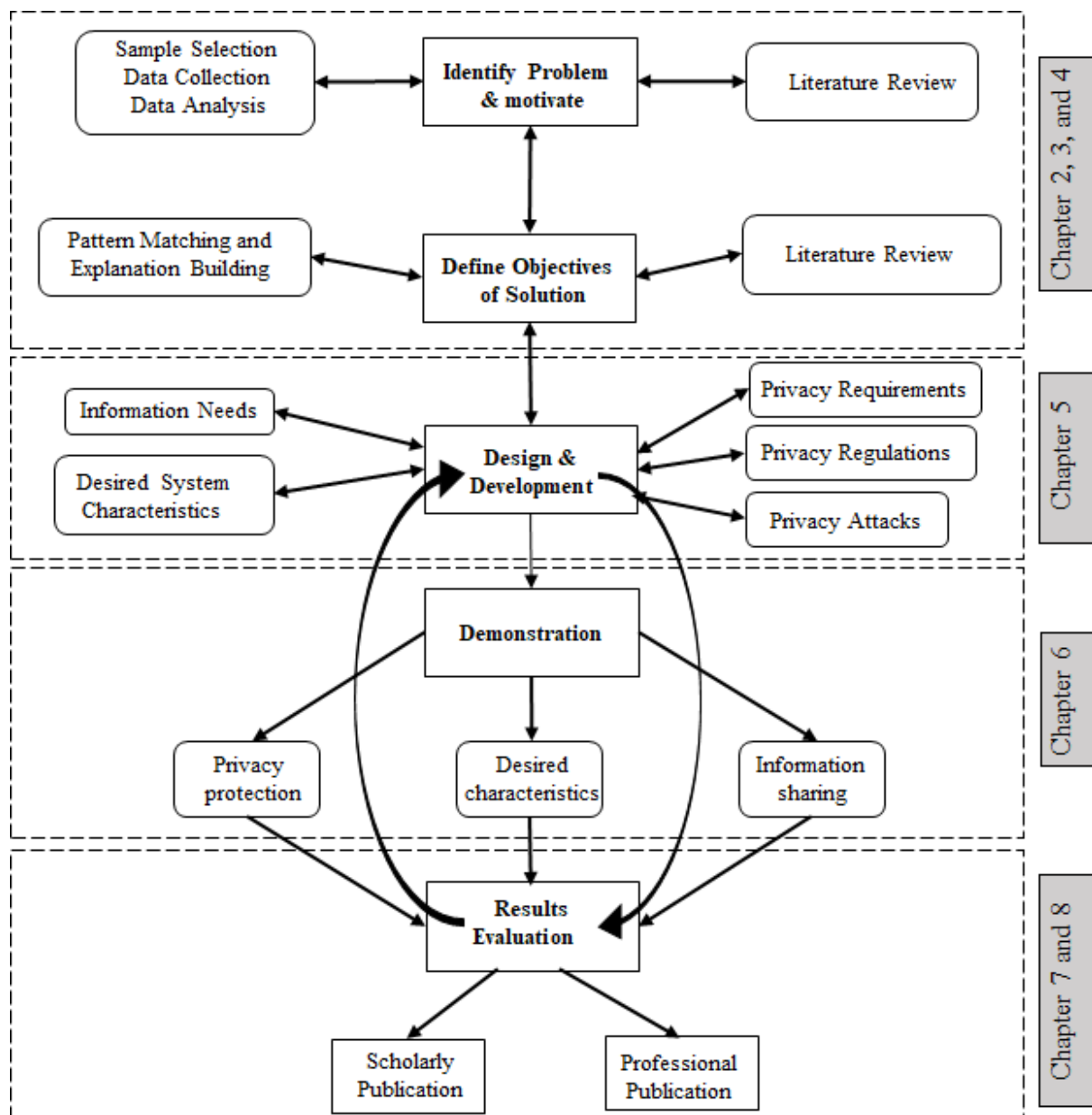


Figure (7.3) Thesis flow in light of the DSRM

After designing and developing the artefact, the **demonstration activity** was performed. The artefact was extensively adapted to the information storing and sharing use. The authors in (Nunamaker, Chen, & Purdin, 1991) stated that when the proposed solution of the research problem cannot be proven mathematically and tested empirically, or if it proposes a new way of doing things, researchers may elect to develop a system to demonstrate the validity of the solution, based on the suggested new methods, techniques, or design. For this research, the researcher has developed scenario-based technical experiments to validate the different aspects involved in the designed artefact which are documented in Chapter 6. The demonstration activity focused on how the artefact enables for sharing healthcare information for medical treatment purposes as well as research purposes in a privacy-preserving manner. Different contexts have been instantiated using randomly generated healthcare datasets. The main goal of the demonstration activity was to assure that the designed artefact meets the objectives of the solution identified in the second activity.

Following the demonstration activity, the **evaluation activity** was conducted. The evaluation activity was conducted based on the model's validity in terms of sharing information in a privacy-preserving manner. The validity of the artefact referred to the substantiation that the artefact, within its applicability in sharing healthcare data in a privacy-preserving manner, possesses a satisfactory range of accuracy consistent with the intended application of it. For this research, the artefact was evaluated using the predictive validation technique described in (Sargent, 2009). In the predictive validation technique, the artefact was used to predict its behaviours. Comparisons then are made between the artefact's behaviors and the prediction made if they are the same. The predictive validation technique was used to compare the output of the designed architecture against the expectations of it. The cloud architecture was expected to meet the objectives of solutions (characteristics) and the evaluation was made based on meeting these objectives.

The last activity was the **communication activity**. This chapter is considered part of the communication activity. The final outcome of this research is a thesis that will be kept at the Auckland University of Technology Library. Moreover, the outcome will also be published in academic journals, academic conference proceedings, and professional outlets.

7.3 Research Questions Evaluation

The main objective of this research was to design a cloud-based architecture for sharing healthcare information in a privacy-preserving manner. However, for that, it was required to answer the following three fundamental questions stated in Chapter 1 which framed this research:

1. How do we maintain the privacy requirements of healthcare data while it is stored on the cloud?
2. What are the characteristics of a privacy-preserving cloud-based architecture for sharing healthcare information?
3. What information can be disclosed for statistical analysis by cloud providers?

Answering the research questions was key to achieving the intended architecture design. The answer to each question was considered in the design and development of the architecture. This section -as part of establishing the research findings- discusses the relations between the answers to the research questions and the characteristics of the designed architecture. The research questions were answered by themes that emerged from the case study data analysis reported in Chapter 4, and information obtained from the relevant literature reported in Chapter 5. The emerged themes from the case study analysis have fed the research with information about how healthcare information systems are used in terms of storing and sharing information, while the information obtained from the relevant literature has fed the research with information related to patients' expectations regarding the privacy of their information, as well as privacy-related requirements and regulations enforced by relevant entities for storing and using patients' information. The rationale between the answers to the research questions and the design of the cloud architecture is discussed.

7.3.1 Research Question 1 (RQ1)

The first research question aimed to understand how the privacy of healthcare information can be maintained while it is stored on the cloud in light of the way it is needed and used. The privacy of information, while it is stored on the cloud, is related to two main aspects namely the privacy of information while it is stored, and the privacy of it when it is disclosed. Figure (7.4) illustrates a breakdown of the first research question. The aspect of information storage is further broken down into two sub-questions: (1) what information to store, and (2) the information's state-of-the-art while stored on the cloud. The disclosure of information also

related to two sub-questions: (1) what information to disclose, and (2) how information is disclosed when it is needed. Therefore, answering this question required gathering and analyzing information about both aspects; storage and disclosure.

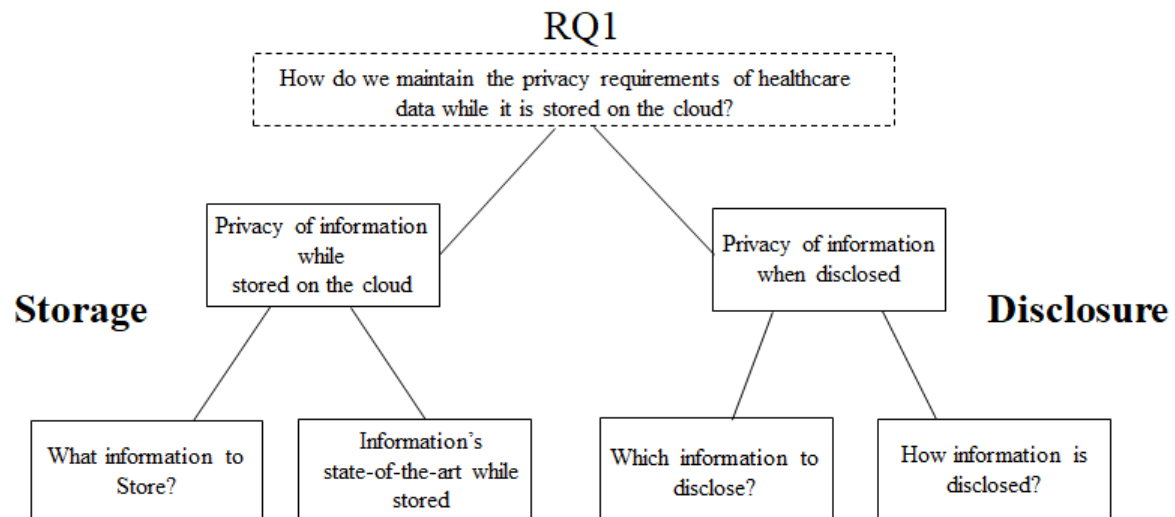


Figure (7.4) Breakdown of the first research question (RQ1)

As seen in the RQ1's breakdown, answering the first research question required answering four questions related to the storage of the information and the disclosure of it. The case study findings reported in Chapter 4 and information obtained from the relevant literature reported in Chapter 2 and Chapter 5 have addressed these questions. Figure (7.4) illustrates the contribution of both case study findings and information from the relevant literature in answering RQ1.

Case study findings

In Chapter 4, the findings of the case study data analysis have given indications of how patients' health information is needed. They also indicated to characteristics that healthcare information systems should have to best serve the healthcare domain. As a result of the data analysis, two themes emerged namely, Information Needs and Desired System Characteristics.

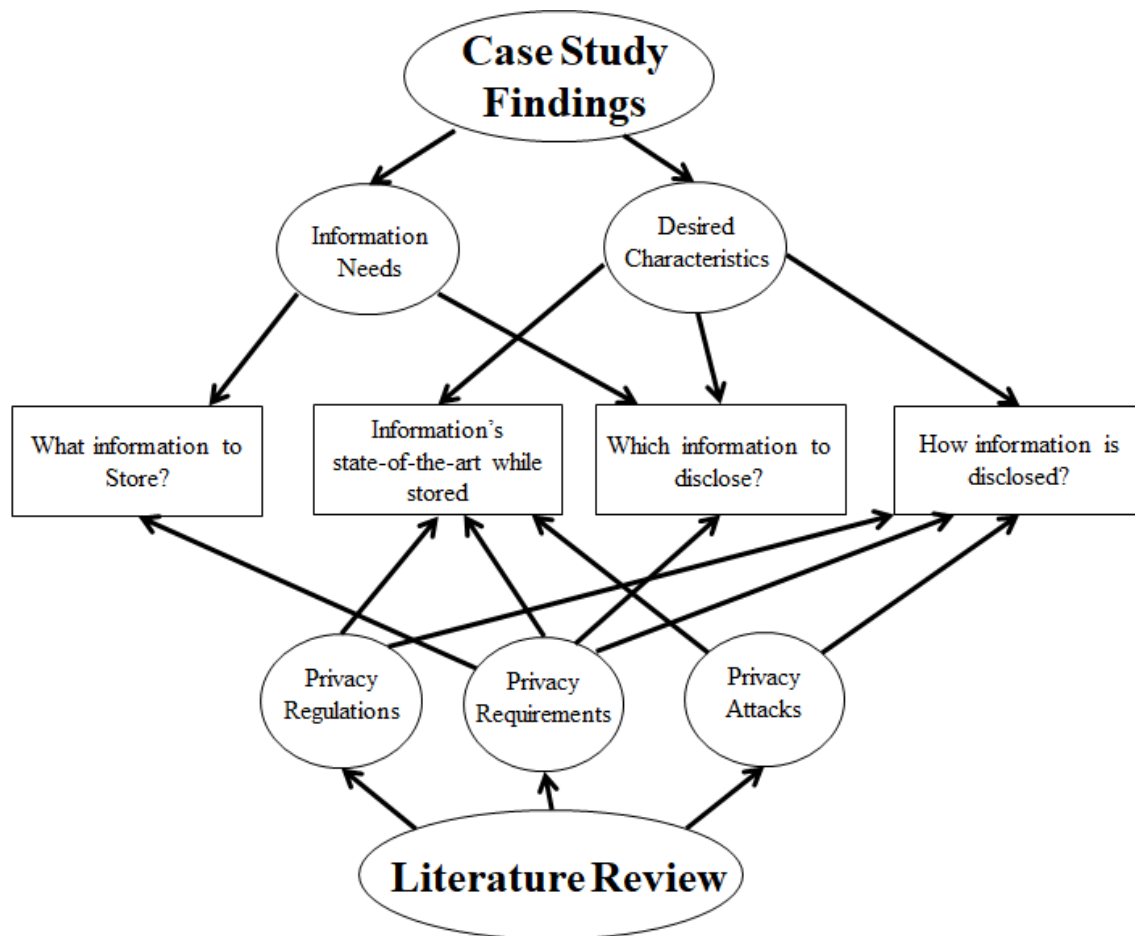


Figure (7.5) Information that answered RQ1

Theme 1: Information Needs

Participants in this study informed the research about the need for patients' health-related information when providing medical treatment to patients. Each of the research participants spoke of ways in which they needed to use and/or share patients' health-related information based on their duties for providing healthcare to patients. All participants explained that all information related to the patients' health is highly important for providing effective medical

treatment to them, however, information related to minor incidents may not be important and therefore storing it permanently may not be useful for future treatment.

Moreover, the analysis of the data collected during interviews has indicated the possibility of categorizing patients' information into 4 different categories. These categories were determined according to the contexts in which patients' information is needed. The derived categories are (1) information that is required for every patients' visit (All_V), (2) information that is required in emergency contexts (Em_V), (3) information that is required in out-patient clinical visits (OutP_V), and finally, (4) information that is required for Research (R). Categorizing patients' information into different categories plays a significant role in protecting the privacy of patients' information, for example, a medical practitioner who works at the emergency department at a hospital may only need to access information that is required in emergencies (Em_V), while a receptionist at the same hospital may only need to access information that is required for all patients' visits (All_V). Therefore, the disclosure of information can be limited to what is needed in each different context. Such findings contributed towards addressing the first research question by answering two of the main sub-questions questions as illustrated in figure (7.4) which are: what information to store and which information to disclose.

Theme 2: Desired System Characteristics

Participants were given a chance to criticize the information systems that they used at work. The goal was to identify how information systems can best serve the healthcare sector in terms of storing and sharing information. Participants in their answers indicated to major challenges that they faced when they used their information systems, these challenges were related to the accessibility of information and its disclosure. Participants in their responses stressed on the need for easy access to the right information, one of the participants said: *"To improve the quality of healthcare provided to patients, the most obvious thing would be is accessing to the right information on the right patient at the right time"*. After reviewing participants' feedback and criticism of the information systems they used, three required characteristics were identified. These characteristics overcome the challenges outlined in participants' responses and make healthcare information systems more effective in terms of storing and sharing healthcare information. These characteristics are related to the storage of information and the disclosure of it.

With regards to the storage of information, storing information in one place was identified as a desired characteristic for ease of access. Not having all patients' information stored in one place creates a problem for medical practitioners to access important information about a patient who is being seen. Patients' information is spread across various locations or holders, therefore, improving the ability to reach the right information about any patient requires locating it. Therefore, storing all patients' information in one place was a characteristic that contributed to answering the question that is related to how information should be stored, or the information's state-of-the-art while stored.

Another identified characteristic in the case study data analysis was about the disclosure of information. Despite the importance of all information about patients, not all information may be required in all instances. Participants stressed the point that disclosing the right information is a key characteristic of healthcare information systems to improve the healthcare services provided to patients. Such characteristics supported the idea of structuring patients' records into the categories identified in the previous theme (Information Needs). In the previous theme, the analysis of the data showed that patients' information can be divided into four categories according to the contexts in which information is needed. In this theme, the analysis showed that Just-efficient information disclosure is a key characteristic that is required in the healthcare information system. Such findings contributed significantly to answering the first research question by addressing the sub-question "which information to disclose?".

Accessing information about patients through a unified platform was also identified as a required characteristic in the healthcare information system. Participants indicated the need for a standardized platform user interface which facilitates easy access to the required information. This identified characteristic was related to the sub-question "How information is disclosed?".

Literature review findings

The information obtained from the relevant literature has contributed significantly to answering the first research question. The findings in the case study analysis have fed the research with information about how healthcare information systems are used for storing and sharing healthcare information, while the information obtained from the relevant literature has provided important information related to the privacy requirements of healthcare information. Information from both sources has given the research important information towards answering the first research question. The relevant literature has given a great deal of information related to privacy regulations, privacy requirements, and privacy attacks. Such information has

completed the image for the researcher in terms of how the privacy requirements of healthcare data should be maintained while stored on the cloud. Information about patients' expectations with regards to the privacy of their information is highly important to consider for maintaining the privacy of information. The privacy-related regulations and policies were essential to consider when answering the first research questions, for example, to assure the privacy of information, it is very important to adhere to legal frameworks such as the Health Insurance Portability and Accountability Act (HIPPA) and the Data Protection Act. Such frameworks clearly specify the responsibilities of organizations with regards to the privacy protection of personal health information. Another important information obtained from the relevant literature was about the privacy attacks that are performed on patient's information while stored and/or disclosed. As part of answering the first research question, it was important to identify potential privacy attacks on healthcare information while stored on the cloud, and consequently, find suitable privacy protection approaches to prevent the information privacy from the identified attacks. Therefore, the answer to the RQ1 was key to achieving the architectural design that is not vulnerable to the identified privacy attacks. It has given the researcher the required knowledge to understand how information that is stored on the cloud is vulnerable to privacy attacks concerning the way it is needed and used.

7.3.2 Research Question 2 (RQ2)

The second research question aimed to identify the characteristics required for designing a privacy-preserving cloud-based architecture to store and share healthcare information

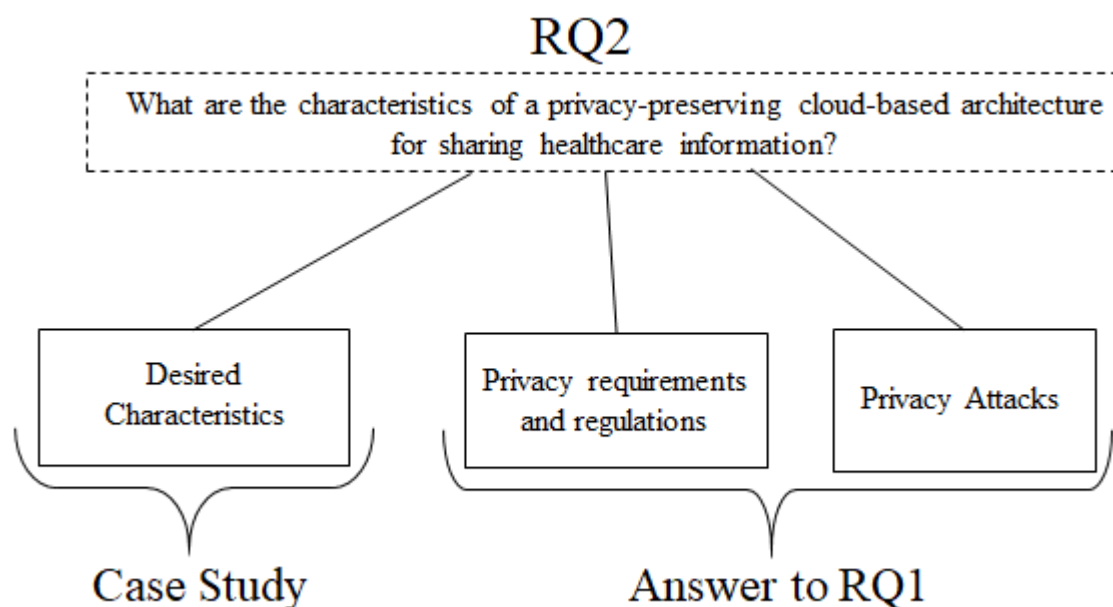


Figure (7.6) Answer breakdown to RQ2

effectively. The answer to the previous research question (RQ1) was a source of information for answering RQ2. In fact, the characteristics required for designing a privacy-preserving cloud-based architecture for sharing healthcare information were derived from the answer to RQ1 and the findings of the case study data analysis.

The answer to RQ1 was related to three main aspects as illustrated in figure (7.7) namely: characteristics that are desired by healthcare information systems' users, privacy-related requirements and regulations for healthcare information, and potential privacy attacks that are identified in the literature. The goal of RQ1 was to understand how the privacy of healthcare information can be maintained while it is stored on the cloud in light of the way it is needed and used. Therefore, the characteristics identified for answering the RQ2 were meant to be

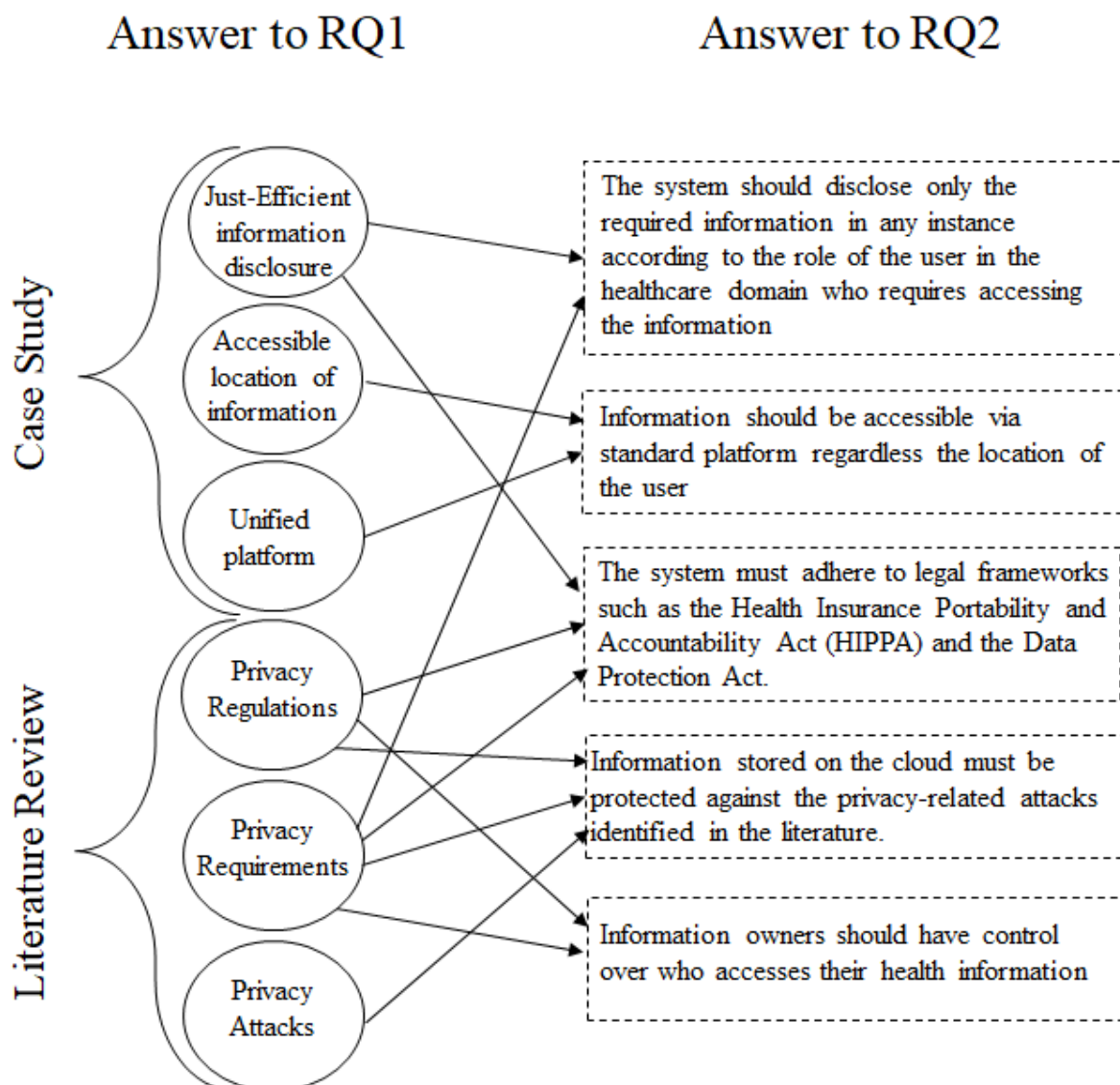


Figure (7.7) Derived system characteristics

solutions to the corresponding requirements and/or desired characteristics identified in the case study findings and relevant literature (answer to RQ1) as illustrated in figure (7.7). As seen in figure (7.7), there were five main characteristics derived from the answer to the RQ1. Each characteristic was proposed as a solution to one or more requirements identified in the case study findings and/or in the relevant literature. These characteristics -in light of the DSRM- were the objectives of solution which were considered in the design of the architecture. Therefore, the answer to the second research question (RQ2) has given the researcher the required information to understand how healthcare information systems can best serve the healthcare domain in terms of storing and accessing information.

7.3.3 Research Question 3 (RQ3)

Sharing healthcare information is important for research purposes, however, there are some challenges related to the privacy of this information when it is shared for research purposes. The main intention of the third research question (RQ3) was to understand how patients' information can be anonymized without affecting the utility of it for research purposes. The goal of the question was to identify pieces of information that could be removed from patients' information for anonymization purposes without affecting the utility of the information for research purposes. The question was answered using information obtained from the case study data analysis and the relevant literature as illustrated in figure (7.8).

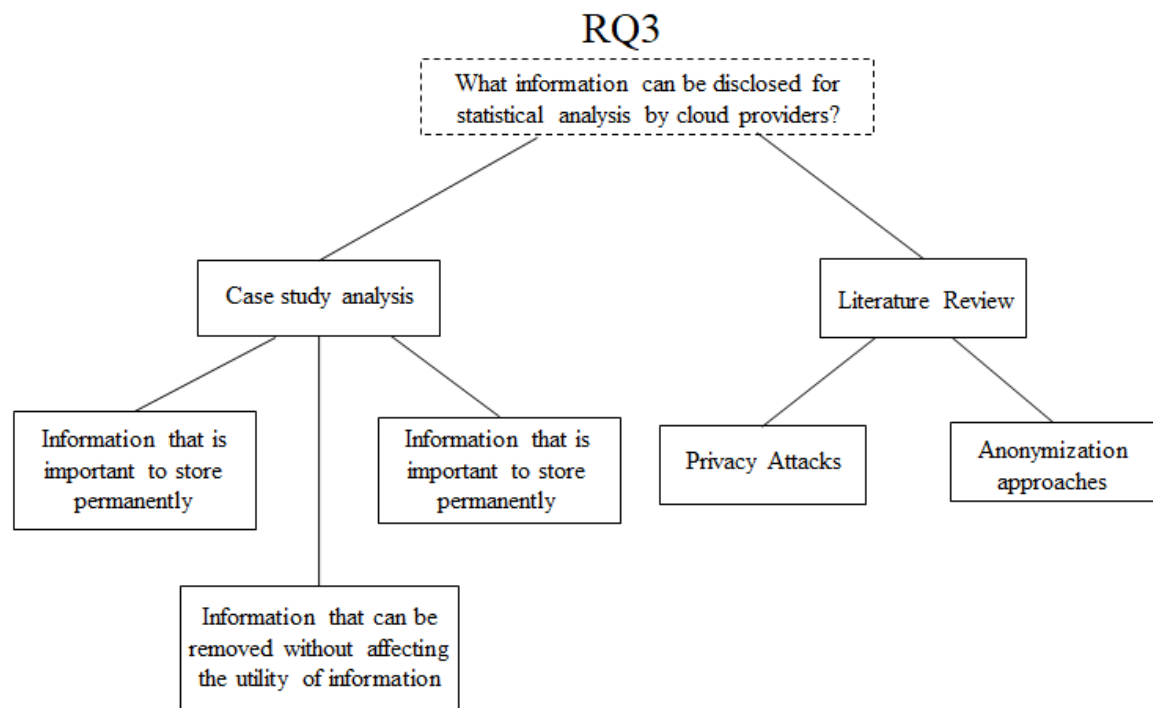


Figure (7.8) RQ3 Answer breakdown

In fact, the anonymity of information is in proportional relationship with the amount of information removed from it, and in an inverse relationship with the utility of it for research purposes, therefore, answering the RQ3 was vital for achieving a trade-off between the anonymity of information when used for research purposes and its accuracy for the same purpose. Such indications have given the researcher knowledge about how anonymization can affect the utility of information when it is used for research purposes.

Case study findings

The case study data analysis has fed the research with information that is significant for answering RQ3. The interview questions included three questions related to the significance of information in terms of storing it for future use. These questions were: “what type of information that is required to store permanently in any particular patient’s health record?”, “what type of information that does not require being stored permanently in any particular patient’s health record?”, and “if researchers or data analysts are to have the information in anonymized form for research purposes, what particular data fields (e.g. Birth Date) can be removed without affecting the outcome of the research?”.

Participants’ answers to these questions have contributed to answering the RQ3 by indicating the information that is important to include in patients’ information when using it for research purposes. Each participant responded to the interview questions as to the way he/she used information when they provided health care services to patients. The analysis of these responses has given the researcher a clear idea about what information is compulsory to have in patients’ information when it is used for research purposes and why. The findings of the case study data analysis have indicated to three types of information in patients’ health-related information; information that can be completely removed without affecting the utility of information for research, information that is important to disclose due to its significance, and finally information that can be disclosed in a generalized form such as age intervals. Such findings have contributed significantly to answering the RQ3 because it gave strong indication and justifications to what information should be included in patients’ information when it is used for research purposes.

Literature review findings

Information obtained from the literature has also contributed significantly to answering the RQ3. The literature has enriched the research with information about how the privacy of individual patients is breached when using their information for research purposes. The

literature has provided significant information about potential privacy attacks that are performed on patients' information when using it for research purposes. The information obtained from the literature has completed the answer to RQ3 by justifying how patients' information may lead to breaching the privacy of individual patients. Such information enabled the researcher to figure out how cloud providers should release patients' information for research purposes concerning the utility and the privacy of it.

7.4 Summary

This chapter has presented an evaluation of the different aspects of the research. The contribution of the research methodology followed toward achieving the outcome artefact (architecture) has been evaluated. The iterations in the activities in light of the DSRM have led to designing an architecture that satisfies the need for sharing healthcare information from a user perspective with consideration to the privacy-related regulations such as HIPPA.

The research questions that have framed this research and the way they were answered have also been justified and evaluated in the chapter. Research questions were answered by analyzing information obtained from both; healthcare practitioners who require using the healthcare information system during their work in the healthcare domain, and the literature. The combination of information sources has enabled to answer the research questions effectively, and this had an impact on the usefulness of the architecture in terms of sharing healthcare information.

Finally, the proposed architecture was evaluated in terms of its ability to share information in a privacy-preserving manner. It was concluded that the proposed architecture satisfies the need for it in the healthcare domain. The implementation of the architecture was evaluated, and its viability and proof of concept were demonstrated and justified.

Chapter 8: Conclusion

Information about patients' health generates special value when it is exchanged and collaboratively used among different parties involved in the healthcare domain. Cloud computing technology appears to be the dreamed vision of the healthcare industry because it matches the need for healthcare information sharing directly to various healthcare-related parties over the internet, regardless of their location and the amount of information being shared. However, the adoption of cloud computing in the healthcare domain has always been hindered due to many challenges in which information privacy is a major one. The purpose of this research was to design a cloud architecture for healthcare information systems to collaboratively share and use information in a privacy-preserving manner. The research was conducted in a multi-methodological approach underpinned by the Design Science research methodology. A case study method was followed for identifying the characteristics required for healthcare information systems. Six healthcare-related institutions participated in the research from which medical practitioners were interviewed.

A cloud architecture design for the healthcare information system was proposed. The proposed architecture enables for storing and sharing patient information for both; medical treatment and research purposes in a privacy-preserving manner. The adoption of the searchable encryption scheme and the separation of information (encrypted data and secret keys) enable the proposed architecture to prevent from the privacy threats that could be performed due to the ability of the cloud provider to read it. The user identity management protocol (U-IDM) preserves the confidentiality of the information that is stored on the cloud and grants patients a means of control over who can access their information. In terms of using patients' information for research purposes, the interactive nature of the proposed architecture eliminates the possibility of successful privacy attacks that could be performed on the dataset when it is released for research purposes. The recursive l -diversity together with controlling the number of tuples and percentage of tuples that contain sensitive target values eliminate the ability of adversaries to identify individual patients or associate a certain value to individuals in any released dataset.

However, to conclude the thesis, this chapter summarizes the research by identifying the key points, research challenges, limitations of the research, and areas for further research. Therefore, this chapter is organized as the following: section 8.1 presents a summary of all key points in the thesis from Chapter one through to this chapter. Section 8.2 outlines and discusses the challenges of the research in both theory and practical perspectives. Section 8.3 presents

the limitation of the research, and finally, section 8.4 discusses the important directions of future research in the field.

8.1 Thesis Summary

As the title of the research suggests, the main goal of the research was designing a privacy-aware cloud-based architecture for sharing healthcare information. Therefore, this research aimed to reach a suitable cloud-based architectural design for the healthcare information systems that satisfies the need for it in terms of storing and collaboratively sharing information in a privacy-preserving manner. This section discusses the scope of this research including a summary of the literature, problem identification, and the proposed cloud architectural design. Moreover, the research methodology and the evaluation method are evaluated in this section accordingly.

Chapter one presented the introduction to the research. The chapter provided a general overview of the issue of sharing healthcare information in terms of the ability of the current healthcare information systems to collaboratively share information, as well as the privacy violation of patients' information when it is shared. Further, the chapter concisely detailed explanation to the research problem and the research questions which have framed this research. Furthermore, the motivation that has triggered the conduction of this research and the significance of it is also outlined in the chapter.

Chapter two presented a number of efforts that have been put by researchers towards enabling the current healthcare information systems to collaborate. The chapter covered two main challenges that hinder such collaboration which are interoperability and privacy. In terms of interoperability, the chapter covered the efforts that have been put by disparate parties towards standardizing medical information for achieving interoperability among healthcare information systems, however, it was concluded that interoperability of electronic information remains a tremendous challenge especially with over 100 electronic healthcare information standards that currently exist and used. There was no existing model found in the literature that is implemented to support the different vocabularies, data interpretation algorithms, and mapping tools in a single source environment; they are all stand-alone applications that hinder interoperability among heterogeneous systems. Data Anonymization is another topic that was covered in the chapter. Aggregating patients' health-related information and sharing it for research purposes is considered highly important in the healthcare domain, however, the privacy of this information must be protected. The chapter explained a number of

anonymization techniques for using patients' health information for research purposes in a privacy-preserving manner. However, the efficiency of these techniques is always at the cost of data utility. Moreover, the chapter explained several privacy-related attacks that could violate the privacy of individual patients when their information is aggregated and shared for research purposes. Nevertheless, the topic of adopting cloud computing technology in the healthcare domain was also presented in detail. Various efforts for protecting the privacy of healthcare information in cloud environments were reviewed.

Chapter three presented the methodological research design followed for the research. The research aimed to propose a solution to the problem of sharing healthcare information in a privacy-preserving manner for medical treatment and research purposes. The research was conducted in a multi-methodological approach underpinned by the Design Science research methodology. A case study method was followed for identifying the characteristics required for healthcare information systems. The chapter outlined and explained the six research activities performed in the research in light of the employed research methodology following the Problem-Centered initiation approach.

Chapter four presented the process of conducting the case study research activities which included: gathering data from research participants, organizing the collected data, and analyzing it. The chapter explained how each activity was conducted. The research participants were recruited from amongst employees of organizations involved in the healthcare sector. 19 individuals from different 6 organizations were recruited for the research. Data was collected through face-to-face interviews with the research participants. Data was organized and analyzed following the steps suggested by Creswell (2007). The chapter also presented and discussed the findings of the data analysis. Two main themes emerged from the data analysis which are information needs and desired system characteristics. These themes have aided the process of identifying the objectives of the solution. The information needs theme has enabled the researcher to categorize patients' information into four categories namely All_V, Em_V, OutP_V, and R, while the desired system characteristics theme enabled the researcher to identify the characteristics that healthcare information systems should have to best serve the healthcare domain.

Chapter five presented the proposed architectural design. The goal of Chapter five was to explain how the proposed architecture is designed in terms of strategies followed and components employed, and further provides details on how they are incorporated towards

sharing information in a privacy-preserving manner. The chapter presented the proposed architecture design in two parts, the first part covered the fundamentals of the design and its comprising components for storing and sharing information for medical treatment purposes. The fundamentals of the design (structuring of patients' health information and using the searchable symmetric encryption scheme) were explained, and their contribution towards enabling the architecture to share information in a privacy-preserving manner was explained. The second part presented the fundamentals and components of the architecture for sharing patients' health information for research purposes. The part first explained the privacy threats on patients' information when used for research purposes which were identified in the literature, and then outlined the strategies followed in the proposed design to protect the information from such threats. Further, the process of releasing patients' information for research purposes is presented with clarification of each activity performed in the process.

Chapter six presented a demonstration of how the proposed architecture stores and shares information in a privacy-preserving manner. The main objective of the chapter was to prove the validity of the proposed architecture in terms of sharing healthcare information in a privacy-preserving manner. The Chapter was organized in two parts and each of them presented a scenario-based instantiation of the different aspects of the proposed architecture. The first part of the chapter presented an instantiation of how the proposed architecture enables for storing and sharing patients' health information for medical treatment purposes without violating the privacy of the information. The second part of the chapter presented an instantiation of how patients' health information is used for research purposes in the proposed architecture. The instantiation illustrated the interactive nature of the proposed architecture in terms of requesting and releasing patients' information for research purposes, with the elimination of the privacy threats outlined in chapter five.

In **Chapter seven**, the contribution of the research methodology to the outcome of the research, the research questions, and the proposed architectural design have been discussed. Section (7.1) presented a reflection on how the Design Science research methodology has contributed towards achieving the outcome of this research, the research activities were outlined and mirrored with the chapters of this thesis. Section (7.2) presented an evaluation of how the research questions were answered as well as their contribution towards achieving the proposed architecture design, while section (7.3) presented an evaluation of the proposed architecture design in terms of its ability to share healthcare information in a privacy-preserving manner. Finally, the research was concluded in **Chapter eight**. The conclusion chapter included a

summary of the entire research, challenges encountered in the research, the limitation of the research, and finally direction for further research in the field.

8.2 Research Challenges

One of the challenges encountered during the conduction of the research was getting access to medical practitioners for the data collection phase. The process of approaching medical practitioners of the invited organizations required obtaining approval from the management of these organizations first. The researcher invited 18 organizations to participate in the research and only 6 organizations have shown a positive attitude towards participating in the research who did not respond promptly to the invitation sent to them. This has caused a significant delay in the process of collecting the data for the research.

Another challenge was related to the ability of the researcher to understand various terms and phrases of the data collected especially during the process of transcription. The researcher had to search online for the meanings of various terms and phrases to understand them during the process of transcription of the interviewee's responses. Research participants responded to the interview questions as to the way they used and shared information for medical care purposes, and it was apparent that they did not consider the researcher's inability to fully understand the meanings of some medical terms and phrases they used in the responses. The researcher spent significant time searching for interpretations and meanings of vocabulary used in the data collected. This challenge has delayed the analysis of the collected data.

Further, after establishing the artefact, the researcher faced challenges related to the implementation of the artefact. As per the research methodology followed for this research, it was required to implement the proposed architecture for performing the demonstration and the activity iterations. However, the implementation of the proposed architecture required programming skills which the researcher did not master. For that, the researcher sought assistance from experts in cloud computing technology and the AWS platform was recommended as a solution to implement the architecture. The researcher had to learn and master a number of services offered by AWS which were used for the implementation of the architecture.

8.3 Research Contribution and Limitations

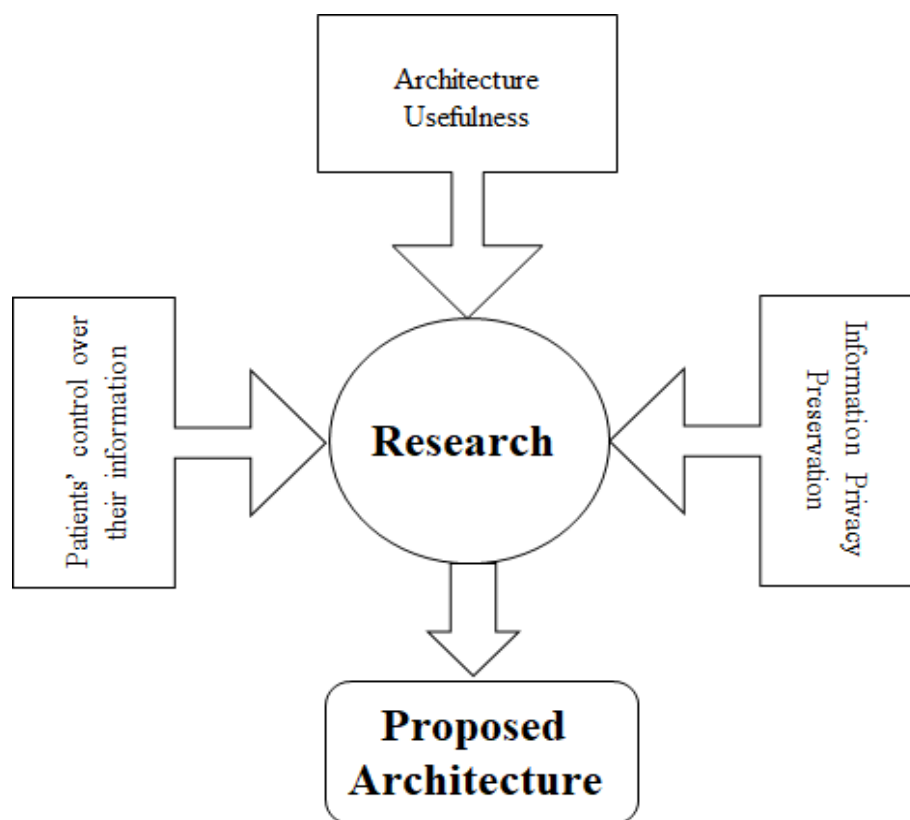


Figure (8.1) Dimensional consideration in the proposed architecture

This research employed rigorous methodology in multiple ways. The research aimed to design a cloud architecture for the healthcare information systems with dimensional consideration to three important aspects which are (1) the usefulness of the architecture in terms of storing and sharing healthcare information, (2) the information's privacy preservation while it is stored and shared using the architecture, and finally (3) the means of control that patients have over who can access their information that is stored on the architecture. For that, the research has been fed with multiple types of data to cover these aspects.

For the usefulness of the architecture, data was collected through interviews with medical practitioners who required storing and using patients' health-related as part of their daily work activities. The analysis of this data has informed the research about how healthcare information systems are used and the information that is required when providing healthcare services to patients or for research purposes. Moreover, the findings of the interview data analysis have indicated to the characteristics that healthcare information systems should have to best serve

the domain. Medical practitioners are considered a reliable source of information because each practitioner responded to interview questions as to the way he/she needed to use information systems during their work in the healthcare sector. The roles of medical practitioners in the healthcare field who participated in the research varied. Nine practitioners were from general practice and urgent care institutions, four of them were from tertiary healthcare institution, three of them were from hospice hospital, one was from urgent and ambulance care, while the last one was a pharmacist. Given the importance of context in qualitative research, this was a strength of the present research. By getting different viewpoints of research participants, it was possible to achieve a greater level of depth in terms of understanding how healthcare information systems can best serve the domain in terms of sharing healthcare information.

The privacy-preservation characteristic of the proposed architecture was achieved by reviewing various data privacy approaches and mechanisms available in the literature. The research was fed with a significant deal of information related to the privacy requirements of healthcare information, privacy regulations of healthcare information such as HIPPA, Privacy Act 1993, Data Protection Act, and privacy-related attacks that could potentially be performed on patients' information. Nevertheless, information obtained from the literature has informed the research about the various efforts that have been made in the field of data privacy and cloud computing data privacy. Collecting and reviewing such information has enabled the researcher to achieve with high confidence a privacy-preserving architectural design that is suitable for the healthcare information systems. Moreover, by reviewing the potential privacy attacks in the literature that can potentially be performed on patients' information, the researcher could achieve a design that is not prone to any of these privacy attacks, and this was another strength of the present research and its outcome.

Patients' control over who can access their health information was another dimension of consideration in the architecture design. The researcher has reviewed the mechanisms related to authentication, authorization, and access control that are available in the literature. Each mechanism was evaluated in terms of its suitability as a solution to control access to information while it is stored on the cloud. Data encryption schemes, access control mechanisms, and identity management approaches were the topics reviewed in the literature for enabling patients to have control over who can access their information. Achieving a design that enables patients to have control over who can access their health information was another strength of the proposed architecture. The comprehensiveness of consideration and the

reliability of the information that fed the research were a strength to the present research in terms of its outcome and its contribution to the body of knowledge.

This research had a limitation as well. Patients' information was categorized into 4 main categories according to the need of it. The number of research participants has caused a limitation to the research. The researcher during the conduction of the research realized the complexity of the healthcare domain in terms of the number of departments and areas in the domain. The researcher acknowledges that the analysis of the data would have led to a bigger number of information categories if more participants from the different areas in the healthcare domain were involved in the research. Moreover, it was noticed that much information may overlap in different instances, therefore, in the real implementation of the proposed system, structuring patients' information can happen in a case by case according to medical conditions that patients may have or the medical treatments that are required. This limitation was due to the limited access to medical practitioners in the various fields and the time allowed for the data analysis phase of the research.

8.4 Directions for Further Research

The adoption of cloud computing for the healthcare information systems is a major improvement in the healthcare domain, due to the significant benefits that cloud computing technology offers. This research has proposed a cloud architecture for the healthcare domain. The feasibility and usability of the proposed architecture were confirmed, and the validity of the architecture in terms of preserving the privacy of information was successfully tested and proven. The research findings and outcomes provide multiple directions for extending and expanding upon the scope and focus of the present research.

Firstly, getting feedback from medical practitioners on the prototype of the designed architecture is an important direction of future research. The proposed architecture has satisfied the need for sharing healthcare information in a privacy-preserving manner, however, getting feedback from the medical practitioners and experts from the healthcare domain may further validate and improve the design of the architecture to best serve the domain.

Secondly, it would be of interest to carry out further investigations related to categorizing patients' information. Patients' information in the present research has been categorized into four categories of which three were for medical treatment purposes, however, a research direction would refine these categories to further limit the exposure of information when it is

needed for medical treatment purposes. This direction would require deeper knowledge in the medical field to allow feeding the research with more technical data related to what and when patients' health information is needed. For example, information about a patient who is diagnosed with diabetes may not be categorized in the same way of categorizing information of a healthy patient, in other words, information categories may contain information that differs from a patient to another according to medical conditions and diseases. For that, it is compulsory to involve experts from the healthcare sector in the research to aid the process of refining the categories of patient information that may be required in different contexts.

Moreover, the proposed architecture allows for manually enrolling patients and storing their information on the cloud, however, healthcare data is today collected using various advanced methods such as mobile devices, wearable sensors, and home wireless networks which can automatically transmit and receive data. Researchers have proven that utilizing the data collected in these methods contributes significantly to healthcare service betterment. Therefore, a research direction can be to expand the proposed architecture design to accommodate patient-generated information that is collected by these data collecting methods.

References

- Abdalla, M., Bellare, M., Catalano, D., Kiltz, E., Kohno, T., Lange, T., . . . Shi, H. (2005). Searchable Encryption Revisited: Consistency Properties, Relation to Anonymous IBE, and Extensions. *Annual International Cryptology Conference* (pp. 205-222). Berlin, Heidelberg: Springer.
- Abdalla, M., Benhamouda, F., & Pointcheval, D. (2016). Public-key encryption indistinguishable under plaintext-checkable attacks. *IET Information Security* (pp. 288 - 303). IET.
- Abdullah, A. M. (2017). Advanced Encryption Standard (AES) Algorithm to Encrypt and Decrypt Data. *Department of Applied Mathematics & Computer Science, Eastern Mediterranean University - Cyprus* .
- Abid, M., Malik, M. S., Usman, M., Hasan, M. M., & Khalid, Z. (2018). A Knn Based Multiple Forms of Attack Prevention Algorithm for Non-Numerical Big Data in Medical Domain. *International Journal of Computer Science and Network Security*, 18(12).
- Adebesin, F., Foster, R., Kotz, P., & Greunen, D. v. (2013). A review of interoperability standards in e-Health and imperatives for their adoption in Africa. *South African Computer Journal*, 55-72.
- Adler-Milstein, J., Sarma, N., Woskie, L. R., & Jha, A. K. (2014). A Comparison Of How Four Countries Use Health IT To Support Care For People With Chronic Conditions. *HEALTH AFFAIRS* 33,, 1559-1566.
- Aggarwal, R. (2018). Resource Provisioning and Resource Allocation in Cloud Computing Environment . *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 1040-1049.
- Ahier, B. (2015, January 06). *FHIR and the future of interoperability*. Retrieved March 01, 2017, from HealthcareITNews: <http://www.healthcareitnews.com/news/fhir-and-future-interoperability>
- Al-Mamun, A., Aseltine, R., & Rajasekaran, S. (2016). Efficient Record Linkage Algorithms Using Complete Linkage Clustering. *PLoS ONE*, 1-21.

- Al-Qurishi, M., Rahman, S. M., Hossain, M. S., Almogren, A., Alrubaian, M., Alamri, A., . . . Gupta, B. (2018). An effecient key agreement protocol for Sybil-precaution in online social networks . *Future Generation Computer Systems*, 139-148.
- Alterovitz, G., & Yao, H. (2015). A Genomics Plan for FHIR. *FHIR Genomics for January 2016 Connectathon* , 1-43.
- Alterovitz, G., Warner, J., Zhang, P., Chen, Y., Ullman-Cullere, M., Isaac, D. K., & Kohane, S. (2015). SMART on FHIR Genomics: facilitating standardized clinico-genomic apps. *American Medical Informatics Association*, 1173-1178.
- Amazon . (2019). *aws*. Retrieved from Amazon Web Services : <https://aws.amazon.com>
- Appari, A., & Johnson, M. E. (1997). Information Security and Privacy in Healthcare: Current State of Research. *Center for Digital Strategies, Tuck School of Business, Dartmouth College, Hanover NH*.
- ASTM. (2012). ASTM International Standards for Healthcare Services, Products and Technology. *ASTM International* .
- Atun, R. (2012). Health systems, systems thinking and innovation. *Health Policy and Planning*, 27:iv4–iv8.
- AWS. (2019). *AWS Documentation*. Retrieved October 02, 2019, from Amazon Web Services: <https://docs.aws.amazon.com/>
- Aziz, H., & Guled, A. (2016). Cloud Computing and Healthcare Services. *Journal of Biosensors & Bioelectronics*.
- Baker, E., Friede, A., Moulton, A., & Ross, D. (1995). CDC's Information Network for Public Health Officials (INPHO): a framework for integrated public health information and practice. *Journal of Public Health Management and Practice : Jphmp*.
- Bandyopadhyay, S., Balamuralidhar, P., & Pal, A. (2013, August). Interoperation among IoT Standards. *Journal of ICT Standardization*, 253–270. doi:10.13052
- Barakat, M., Eder, C., & Hanke, T. (2018). An Introduction to Cryptography. *the University of Kaiserslautern*.
- Bazeley, P., & Jackson, K. (2013). *Qualitative Data Analysis with NVivo*. UK: SAGE.

- Begoyan, A. (2007). An Overview of Interoperability Standards for Electronic Health Records. *Integrated Design and Process Technology*, 1-8.
- Bélanger, F., & Crossler, R. E. (2011, December). Privacy in the digital age: a review of information privacy research in information systems. *MIS Quarterly*, 35(4), 1017-1042.
- Bellare, M., Boldyreva, A., & O'Neill, A. (2007). Deterministic and Efficiently Searchable Encryption. *Annual International Cryptology Conference* (pp. 535–552). Berlin, Heidelberg: Springer .
- Belle, A., Thiagarajan, R., Soroushmehr, S. M., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big Data Analytics in Healthcare. *BioMed Research International*, 1-17.
- Benaloh, J., Chase, M., Horvitz, E., & Lauter, K. (2009). Patient Controlled Encryption: Ensuring Privacy of Electronic Medical Records. *Microsoft Research, Redmond, WA, USA*.
- Benson, T. (2012). *Principles of health interoperability HL7 and SNOMED*. Springer.
- Berg, B. L. (2004). *Qualitative research methods for the social sciences (5th ed.)*. Boston: Pearson.
- Bertino, E., Bonatti, P. A., & Ferrari, E. (2001). TRBAC: A temporal role-based access control model. *CM Transactions on Information and System Security (TISSEC)*, 191-233.
- Bertino, E., Bonatti, P. A., & Ferrari, E. (2001). TRBAC: A temporal role-based access control model. *ACM Transactions on Information and System Security (TISSEC)*, 191-233.
- Bethencourt, J., Sahai, A., & Waters, B. (2007). Ciphertext-policy attribute based encryption . *Symp. Secur. Privacy*, (pp. 321-334). IEEE.
- Blackman, S. M. (2017). Towards a Conceptual Framework for Persistent Use: A Technical Plan to Achieve Semantic Interoperability within Electronic Health Record Systems. *Proceedings of the 50th Hawaii International Conference on System Sciences*, (pp. 4653-4662).
- Blaze, M., Bleumer, G., & Strauss, M. (1998). Divertible protocols and atomic proxy cryptography. In Nyberg K. (eds) *Advances in Cryptology — EUROCRYPT'98* (pp. 127-144). Berlin, Heidelberg: Springer.
- Blythe, J., & Royle, J. A. (1993). Assessing nurses' information needs in the work environment. *Bulletin of the Medical Library Association*, 81(4), 433-435.

- Bock, C. E., Carnahan, L. J., Fenves, S. J., Gruninger, M., Kashyap, V., Lide, B. B., . . . Sriram, R. D. (2005). Healthcare Strategic Focus Area: Clinical Informatics. *National Institute of Standards and Technology, Technology Administration*, 1-33.
- Boneh, D., Crescenzo, G. D., Ostrovsky, R., & Persiano, G. (2004). Public key encryption with keyword search. *International Conference on the Theory and Applications of Cryptographic Techniques* (pp. 506-522). Berlin, Heidelberg: Springer.
- Boneh, D., Crescenzo, G. D., Ostrovsky, R., & Persiano, G. (2004). Public Key Encryption with keyword Search. *International Conference on the Theory and Applications of Cryptographic Techniques* (pp. 506-522). Berlin, Heidelberg: Springer.
- Borgohain, T., Kumar, U., & Sanyal, S. (2015). Survey of Security and Privacy Issues of Internet of Things. *International Journal of Advanced Networking and Applications*, 6(4), 2372-2379.
- Bowles, K. H., Potashnik, S., Ratcliffe, S. J., Rosenberg, M., Shih, N.-W., Topaz, M., . . . Naylor, M. D. (2013). Conducting research using the electronic health record across multi-hospital systems: Semantic harmonization implications for administrators. *Journal of Nursing Administration*, 355-360.
- Bradshaw, D., Folco, G., Cattaneo, G., & Kolding, M. (2012). *Quantitative Estimates of the Demand for Cloud Computing in Europe and the Likely Barriers to Uptake*.
- Brailer, D. J. (2005). Interoperability: The Key To The Future Health Care System. *Health Affairs*.
- Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2008). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Elsevier: Future Generation Computer Systems*, 599-616.
- Cantwell, E., & McDermott, K. (2016). Making technology talk: how interoperability can improve care, drive efficiency, and reduce waste. *Healthcare Financial Management*, 70.
- Casola, V., Castiglione, A., Choo, K.-K. R., & Esposito, C. (2016). Healthcare-Related Data in the Cloud: Challenges and Opportunities. *IEEE Cloud Computing*.
- Cavaye, A. (1996). "Case Study Research: a Multi-faceted Research Approach for IS., *Information System Journal* , 227-242.

CDISC . (2013). Retrieved February 23, 2017, from Clinical Data Interchange Standards Consortium.: <http://goo.gl/Wt7HN>

CDISC. (2013). *CDISC: FAQ*. Retrieved February 23, 2017, from Clinical Data Interchange Standards Consortium: <http://goo.gl/yWkkz>

CEN. (2009). *Health Informatics, Published Standards*. Retrieved February 23, 2017, from European Committee for Standardization: <http://goo.gl/MMXY3>

CEN. (2012, December). *Hands on Standardization, A starter Guide to Standardization For Experts in CEN Technical Bodies*. Retrieved February 23, 2017, from European Committee for Standardization :
<ftp://ftp.cen.eu/CEN/Services/Education/Handsonguides/Handsonstandards.pdf>

Chang, Y.-C. (2004). Single Database Private Information Retrieval with Logarithmic Communication. *The 9th Australasian Conference on Information Security and Privacy* (pp. 50-61). Sydney, Australia: Springer-Verlag.

Chauhan, K. K., Sanger, A. K., & Verma, A. (2015). Homomorphic Encryption for Data Security in Cloud Computing. *International Conference on Information Technology (ICIT)*. IEEE.

Chen, D., & Zhao, H. (2012). Data Security and Privacy Protection Issues in Cloud Computing. *International Conference on Computer Science and Electronics Engineering (ICCSEE)*, . IEEE.

Chenthara, S., Ahmed, K., Wang, H., & Whittaker, F. (2019). Security and Privacy-Preserving Challenges of e-Health Solutions in Cloud Computing. *School of Engineering and Science, Victoria University, Melbourne*.

Christodoulakis, C., Asgarian, A., & Easterbrook, S. (2017). Barriers to Adoption of Information Technology in Healthcare. *ACM CASCON conference*. Toronto, Canada.

Claret, O. A. (2011). Overview of Cryptography. *SSRN Electronic Journal*.

Cloud Security Alliance. (2011). *Security guidance for critical areas of focus in cloud computing*. USA: CSA (Cloud Security Alliance). Retrieved from: <https://cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf>

Compton, M., & Mickelberg, K. (2014, October). Connecting Cybersecurity with the Internet of Things. *PricewaterhouseCoopers*.

Corbin, J., & Strauss, A. (1994). *Basics of qualitative research: Techniques and procedures for developing grounded theory (2nd ed)*. Thousand Oak, CA: Sage.

Corbin, J., & Strauss, A. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory (2nd ed)*. Thousand Oaks, CA: Sage.

Creswell, J. (2007). *Qualitative inquiry and research method: Choosing among five approaches (2nd ed.)*. Thousand Oaks, CA: Sage.

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Creswell, J. W. (2007). *Qualitative inquiry and research method: Choosing among five approaches (2nd ed.)*. Thousand Oaks, CA: Sage.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches (3rd ed.)*. Los Angeles: Sage.

Culnan, M. (1993). How Did They Get My Name? *An Exploratory Investigation of Consumer Attitudes Towards Secondary Information Use*, 3(17).

Curtmola, R., Garay, J., Kamara, S., & Ostrovsky, R. (2006). Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions. *13th ACM Conference on Computer and Communications Security*, (pp. 79-88).

Deering, M. J. (2013). Issue Brief: Patient-Generated Health Data and Health IT. *The Office of the National Coordinator for Health Information Technology*, 1-11.

Denecke, K., Bamidis, P., Bond, C., Gabarron, E., Househ, M., Lau, A. Y., . . . Hansen, M. (2015, August). Ethical Issues of Social Media Usage in Healthcare. *Yearb Med Inform*, 10(1), 137–147. doi:10.15265/IY-2015-001

Dialogic. (2010). Introduction to Cloud Computing. *Dialogic*.

Diaz, B. (2016, December 10). *Health Language Blog: What is Semantic Interoperability?* Retrieved February 25, 2017, from Health Language: <http://blog.healthlanguage.com/what-is-semantic-interoperability>

- DICOM. (2011). Part 1: Introduction and Overview . *Digital Imaging and Communications in Medicine (DICOM)* . Retrieved from Digital Imaging and Communications in Medicine.
- Doukas, C., Pliakas, T., & Maglogiannis, I. (2010). Mobile healthcare information management utilizing Cloud Computing and Android OS. *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE.
- Dubey, N., & Vishwakarma, S. (2016). Cloud Computing in Healthcare. *International Journal of Current Trends in Engineering & Research (IJCTER)*, 211-216.
- El-Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., . . . Bottomley, J. (2009). A globally optimal k-anonymity method for the de-identification of health data. *J Am Med Inform Assoc.*, 670-682.
- Elisa Bertino, F. P. (2009). Privacy-preserving digital identity management for cloud computing. *IEEE Data Eng. Bull.*, 21-27.
- Eludiora, S., Abiona, O., Oluwatope, A., Oluwaranti, A., Onime, C., & Kehinde, L. (2011). A User Identity Management Protocol for Cloud Computing Paradigm. *Int. J. Communications, Network and System Sciences*, 4, 152-163.
- Esterberg, K. G. (2002). *Qualitative Methods in Social Research*. McGraw-Hill.
- Etzioni, A. (2010). Personal Health Records Why Good Ideas Sometimes Languish . *Issues in science and technology*, 59-66.
- Ezea, B., & Peyton, L. (2015). Systematic Literature Review on the Anonymization of High Dimensional Streaming Datasets for Health Data Sharing. *Procedia Computer Science*, 63, 348 – 355 . doi: 10.1016/j.procs.2015.08.353
- Fabiana, B., Ermakovab, T., & Junghannsa, P. (2015, March). Collaborative and secure sharing of healthcare data in multi-clouds. *Information Systems*, 48, 132-150. doi:10.1016/j.is.2014.05.004
- Ferguson, J., Hannigan, A., & Stack, A. (2018). A new computationally efficient algorithm for record linkage with field dependency and missing data imputation. *International Journal on Medical Informatics*, 70-75.

Fischer, E. A., & Figliola, P. M. (2013). Overview and Issues for Implementation of the Federal Cloud Computing Initiative: Implications for Federal Information Technology Reform Management. *Congressional Research Service*.

Franz, B., Schuler, A., & Krauss, O. (2015). Applying FHIR in an Integrated Health Monitoring System. *European Journal for Biomedical Informatics (EJBI)*, 11(2).

Frost, J. (2008). Combining approaches to qualitative data analysis: Synthesising the mechanical (CAQDAS) with the thematic (a voice-centred relational approach). *Methodological Innovations Online* , 25-37.

Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010, June). Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Computing Surveys*, 42(4).

Gaboury, I., Bujold, M., Boon, H., & Moher, D. (2009). Interprofessional collaboration within Canadian integrative healthcare clinics: Key components. *Social Science & Medicine*, 69(5), 707–715. doi:10.1016/j.socscimed.2009.05.048

Gabriel, M. H., Furukawa, M. F., Jones, E. B., King, J., & Samy, L. K. (2014). Progress and challenges: Implementation and use of Electronic Health Records among Critical Access Hospitals. *Health Affairs*, 1262-1270.

Gkoulalas-Divanis, A., & Loukides, G. (2015). Introduction to Medical Data Privacy. In *Medical Data Privacy Handbook* (pp. 1-14). Switzerland: Springer International Publishing. doi:10.1007/978-3-319-23633-9

Gkoulalas-Divanis, A., Loukides, G., & Sun, J. (2014). Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of Biomedical Informatics*, 4-19.

Goh, E. (2003). *Secure Indexes* . IACR ePrint Cryptography Archive.

Gowda, S. (2016). Using Blowfish Encryption To Enhance Security Feature Of An Image . *Department of Computer Science And Engineering, R.V.College Of Engineering, Bangalore, India*.

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly* , 611-642.

Griebel, L., Prokosch, H.-U., Köpcke, F., Toddenroth, D., Christoph, J., Ines Leb, I. E., & Sedlmayr, M. (2015). A scoping review of cloud computing in healthcare. *BMC Medical Informatics and Decision Making*, 1-16.

Griffith, E. (2016, May 03). *What Is Cloud Computing?* Retrieved March 21, 2017, from PCMag: <http://au.pcmag.com/networking-communications-software-products/29902/feature/what-is-cloud-computing>

Gross, G. (2005, January 10). *"Lack of standards hinders electronic health records. Interoperability concerns loom large"*. Retrieved February 12, 2017, from IDGNS: <http://www.infoworld.com/article/2668312/security/lack-of-standards-hinders-electronic-health-records.html>

Grossman, J. M., Zayas-Cabán, T., & Kemper, N. (2009). Information Gap: Can Health Insurer Personal Health Records Meet Patients' And Physicians' Needs? *Health Affairs*, 28(2), 377-389.

Gunasekara, G., & Dillon, E. (2008). Data Protection Litigation in New Zealand: Processes and Outcomes . *Victoria University of Wellington Law Review (VUWLR)* , 39.

Guo, Y., Kuo, M.-H., & Sahama, T. (2012). Cloud computing for healthcare research information sharing. *IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom)*. IEEE.

Gupta, B., & Quamara, M. (2018). An identity based access control and mutual authentication framework for distributed cloud computing services in IoT environment using smart cards. *Procedia Computer Science* , 189-197.

Hashizume, K., Rosado, D. G., Fernández-Medina, E., & Fernandez, E. B. (2013). An analysis of security issues for cloud computing. *Journal of Internet Services and Applications*.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research . *MIS Quarterly* , 28(1), 75-105.

Hibbard, J. H., & Greene, J. (2013). What The Evidence Shows About Patient Activation: Better Health Outcomes And Care Experiences; Fewer Data On Costs. *Health Affairs*, 32(2), 207-214.

HIMSS. (2016). Playing with FHIR . *InterSystems*.

HL7. (2015). *Introduction to HL7 Standards*. Retrieved February 26, 2017, from Health Level Seven: Health Level Seven, <http://hl7.org/>

HL7. (2016). *Health Level Seven Fast Healthcare Interoperability Resources*. Retrieved February 26, 2017, from Health Level Seven Fast Healthcare Interoperability : <http://www.hl7.org/implement/standards/fhir/>

HL7.(2016).*Home*. Retrieved February 26, 2017, from FHIR: <http://www.hl7.org/implement/standards/fhir/>

HL7. (2017). *HL7 Backgrounder Brief*. Retrieved February 23, 2017, from Health Level Seven: <http://www.hl7.org/newsroom/HL7backgrounderbrief.cfm>

HL7. (2017). *Home Page*. Retrieved February 23, 2017, Retrieved from Health Level Seven: <http://www.hl7.org/>

Hripcsak, G., Bloomrosen, M., FlatleyBrennan, P., Chute, C. G., Cimino, J., Detmer, D. E., . . . Wilcox, A. (2014). Health data use, stewardship, and governance: ongoing gaps and challenges: a report from AMIA's 2012 Health Policy Meeting. *Journal of the American Medical Informatics Association*, 21(2), 204–211. doi:10.1136/amiajnl-2013-002117

Hu, V. C., Ferraiolo, D. F., & Kuhn, D. R. (2006). Assessment of Access Control Systems. *National Institute of Standards Technoly (NAT)*, Gaithersburg, MD, USA.

Hu, V. C., Kuhn, D. R., & Ferraiolo, D. F. (2015). Attribute-Based Access Control. *National Institute of Standards and Technology* , 85-88.

Hu, Y., & Bai, G. (2014). A Systematic Literature Review Of Cloud Computing in Ehealth . *Health Informatics-An International Journal (HIJ)*.

Hussien, A.-e.-e. A., Hamza, N., & Hefny, H. A. (2013, April). Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing. *Journal of Information Security*, 4, 101-112.

IEEE. (2017). *About IEEE*. Retrieved February 23, 2017, from The Institute of Electrical and Electronics Engineers : <http://www.ieee.org/about/index.html>

IEEE. (2017). *Formal Liaisons*. Retrieved February 23, 2017, from Institute of Electrical and Electronics Engineers (IEEE): <http://goo.gl/DLZl8>

IEEE. (2017). *Healthcare IT standards*. Retrieved February 23, 2017, from Institute of Electrical and Electronics Engineers (IEEE): <http://goo.gl/ahz3u>

IHE. (2016). *About IHE*. Retrieved February 23, 2017, from Integrating the Health Enterprise: http://www.ihe.net/About_IHE/

Iroju, O., Soriyan, A., Gambo, I., & Olaleke, J. (2013). Interoperability in Healthcare: Benefits, Challenges and Resolutions . *International Journal of Innovation and Applied Studies*, 262-270.

Ishigure, Y. (2017). *Trends, Standardization, and Interoperability of Healthcare Information*. Retrieved February 20, 2017, from NTT Technical Review, Retrieved from <https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201104gls.html>

Islam, S. M., Kwak, D., Kabir, M. H., Hossain, M., & Kwak, K.-S. (2015). The Internet of Things for Health Care: A Comprehensive Survey. *IEEE Access*, 3, 678 - 708.

ISO. (2004). Standardization and Related Activities - General Vocabulary. *ISO/IEC Guide 2*.

ISO. (2013). *Business Plan*. Retrieved February 21, 2017, from Business Plan: ISO/TC 215 Health Informatics.

ISO. (2017). *About ISO*. Retrieved February 21, 2017, from ISO: <http://www.iso.org/iso/home/about.htm>

ISO. (2017). *ISO Membership Manual*. Retrieved February 22, 2017, from ISO: http://www.iso.org/iso/iso_membership_manual.pdf

ITU. (2011). *Standards and eHealth*. Retrieved February 15, 2017, from ITU: https://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000120003PDFE.pdf

Jabbar, I., & Najim, S. (2016). Using Fully Homomorphic Encryption to Secure Cloud Computing. *Internet of Things and Cloud Computing*, 13-18.

Jain, J., & Singh, A. (2017). A Survey on Security Challenges of Healthcare Analysis Over Cloud. *International Journal of Engineering Research & Technology (IJERT)*, 905-912.

Jamshed, S. (2014). Qualitative research method-interviewing and observation. *Journal of Basic and Clinical Pharmacy* , 87-88.

Jiang, S., Cao, Y., Iyengar, S., Kuryloski, P., Jafari, R., Xue, Y., . . . Wicker, S. (2008). CareNet: an integrated wireless sensor networking environment for remote healthcare. *BodyNets '08*

Proceedings of the ICST 3rd international conference on Body area networks. Brussels, Belgium.

John, W. S., & Johnson, P. (2004). The Pros and Cons of Data Analysis Software for Qualitative Research. *Journal of Nursing Scholarship* .

Kagadis, G. C., Kloukinas, C., Moore, K., Philbin, J., Papadimitroulas, P., Alexakos, C., . . . Hendee, W. R. (2013). Cloud computing in medical imaging. *The International Journal of Medical Physics Research and Practice* .

Kamara, S. (2010). Cryptographic cloud storage. *International Conference on Financial Cryptography and Data Security*, (pp. 136-149).

Karla Felix Navarro, E. L., & Lim, B. (2009). Medical MoteCare: A Distributed Personal Healthcare Monitoring System. *International Conference on eHealth, Telemedicine, and Social Medicine* (pp. 25 - 30). Cancun: IEEE. doi:10.1109/eTELEMED.2009.19

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. *Sixth International Conference on Contemporary Computing (IC3), IEEE*, 404–409.

Kepes, B. (2011). Understanding the cloud computing stack: SaaS, PaaS, IaaS. *Rackspace Hosting, Diversity* , 1-17.

Khalilia, M., Choi, M., Henderson, A., Iyengar, S., Braunstein, M., & Sun, J. (2015). Clinical Predictive Modeling Development and Deployment through FHIR Web Services. *American Medical Informatics Association*, 717–726.

Khan, M. A. (2016). A survey of security issues for cloud computing. *Journal of Network and Computer Applications*, 11-29.

Khan, W. A., Khattak, A. M., Hussain, M., Amin, M. B., Afzal, M., Nugent, C., & Lee, S. (2014). An Adaptive Semantic based Mediation System for Data Interoperability among Health Information Systems. *Journal of Medical Systems*.

Kiran, P., & Kavya, N. P. (2012). A Survey on Methods, Attacks and Metric for Privacy Preserving Data Publishing. *International Journal of Computer Applications*, 20-28.

Kiryakova, G., Angelova, N., & Yordanova, L. (2015). Application of Cloud Computing Services in Business . *Trakia Journal of Sciences*, 392-396.

- Kitamura, T., Kiyohara, K., Matsuyama, T., Hatakeyama, T., Shimamoto, T., Izawa, J., . . . Iwami, T. (2016, March 5). Is Survival After Out-of-Hospital Cardiac Arrests Worse During Days of National Academic Meetings in Japan? A Population-Based Study. *Journal of Epidemiology*, 26(3), 155-162. doi:10.2188/jea.JE20150100
- Kokkinaki, A., Chouvarda, I., & Maglaveras, N. (2006). Integrating SCP-ECG files and patient records: an ontology based approach. Greece: University of Thessaloniki.
- Kumar, P., & Lee, H.-J. (2012). Security Issues in Healthcare Applications Using Wireless Medical Sensor Networks: A Survey. *Sensors* 2012, 12(1), 55-91. Retrieved from <http://www.mdpi.com/1424-8220/12/1/55/htm>
- Kuo, A. M.-H. (2011). Opportunities and Challenges of Cloud Computing to Improve Health Care Services. *J Med Internet Res*.
- Kuo, M.-H., Lai, F., Dorjgochoo, S., & Jigjidsuren, C. (2012). A Cloud Computing Based Platform for Sharing Healthcare Research Information. *4th International Conference on Cloud Computing Technology and Science (CloudCom)* (pp. 504-508). IEEE.
- Kuribayashi, S.-i. (2012). Reducing Total Power Consumption Method in Cloud Computing Environments. *International Journal of Computer Networks & Communications*, 69-84.
- Kuzel, A. J. (1999). Sampling in Qualitative Inquiry. In B. F. Miller, *Doing Qualitative Research* (pp. 33-45). Thousand Oaks: Sage Publications.
- Lau, L. M., & Shakib, S. (2005). Towards Data Interoperability: Practical Issues in Terminology Implementation and Mapping. *77th AHIMA Convention and Exhibit*.
- Lazar, D., Chen, H., Wang, X., & Zeldovich, N. (2014). New Number-Theoretic Cryptographic Primitives. *APSys* (pp. 1-7). Beijing, China : ACM.
- Lee, C. H., Kim, Y. S., & Lee, Y. H. (n.d.). Implementation of SMART APP Service Using HL7_FHIR.
- Lester, M., Boateng, S., Studeny, J., & Coustasse, A. (2016). Personal Health Records: Beneficial or Burdensome for Patients and Healthcare Providers? *Perspectives in Health Information Management*.

- Li, M., Yu, S., Zheng, Y., & Ren, K. (2013). Scalable and Secure Sharing of Personal Health Records in Cloud Computing Using Attribute-Based Encryption. *Transactions on Parallel and Distributed Systems* (pp. 131-143). IEEE.
- Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-Anonymity and I-Diversity. *Center for Education and Research, Information Assurance and Security*.
- Li, Y., Bai, C., & Reddy, C. K. (2016, February 10). A distributed ensemble approach for mining healthcare data under privacy constraints. *Information Sciences*, 330, 245–259.
- Li, Z.-R., Chang, E.-C., Huang, K.-H., & Lai, F. (2011). A Secure Electronic Medical Record Sharing Mechanism in the Cloud Computing Platform . *15th International Symposium on Consumer Electronics* (pp. 98-103). IEEE.
- Liu, Y., Sun, Y. (., Ryoo, J., Rizvi, S., & Vasilakos, A. V. (2015). A Survey of Security and Privacy Challenges in Cloud Computing: Solutions and Future Directions. *Journal of Computing Science and Engineering*, 119-133.
- Loukides, G., & Shao, J. (2011). Preventing range disclosure in k-anonymised data. *Expert Systems with Applications*, 4559-4574.
- Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). l-Diversity: Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data*. Atlanta, GA, USA .
- Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). ℓ -Diversity: Privacy Beyond k-Anonymity. *Proceedings of the 22nd International Conference on Data Engineering* (p. 24). Atlanta: IEEE. doi:10.1109/ICDE.2006.1
- Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2007). l-diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 3.
- Maheshwarkar, N., Pathak, K., & Choudhari, N. S. (2012). K-anonymity Model for Multiple Sensitive Attributes. *Special Issue of International Journal of Computer Applications on Optimization and On-chip Communication*, 51-56.
- Maheu, M., Whitten, P., & Allen, A. (2001). *E-Health, Telehealth, and Telemedicine: A Guide to Startup and Success*. Jossey-Bass: Wiley.

Mamlin, B. W., & Tierney, W. M. (2016, January). The Promise of Information and Communication Technology in Healthcare: Extracting Value From the Chaos. *The American Journal of The Medical Sciences*, 351(1), 59-68.

Manta, A. (2013, November). Literature Survey on Privacy Preserving Mechanisms for Data Publishing. *Department of Intelligence Systems, Faculty EEMCS, Delft University of Technology*.

McLellan, E., MacQueen, K. M., & Neidig, J. L. (2003). Beyond the Qualitative Interview: Data Preparation and Transcription. *Field Methods*, 63-84.

Mehraeen, E., Ghazisaeedi, M., Farzi, J., & Mirshekari, S. (2017). Security Challenges in Healthcare Cloud Computing: A Systematic Review . *Global Journal of Health Science*, 137-166.

Meingast, M., Roosta, T., & Sastry, S. (2006). Security and privacy issues with health care information technology. In Engineering in Medicine and Biology Society. *28th Annual International Conference of the IEEE* (pp. 5453-5458). IEEE.

Mell, P., & Grance, T. (2011). The NIST Definition of Cloud Computing. *National Institute of Standards and Technology* . NIST.

Mell, P., & Timothy, G. (2011). *The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology*.

Microsoft. (2019). *Microsoft SQL Server*. Retrieved from Microsoft SQL Server. <https://www.microsoft.com/en-us/sqlserver/default.aspx>

Mishra, M., Das, A., Kulkarni, P., & Sahoo, A. (2012). Dynamic Resource Management Using Virtual Machine Migrations. *Cloud Computing: Networking and communication challenges*. IEEE Communication Magazine.

Mohamed, R., & Harb, H. M. (2015). Public-Key Cryptography Techniques Evaluati. *International Journal of Computer Networks and Applications*.

Montgomery, K., Mundt, C., Thonier, G., Tellier, A., Udoh, U., Barker, V., . . . Kovacs, G. (2004). Lifeguard - a personal physiological monitor for extreme environments. *Engineering in Medicine and Biology Society, 2004. IEMBS '04. 26th Annual International Conference of the IEEE. 1*. San Francisco: IEEE.

- Natsuki, T. (2008). Trends in Healthcare Information Standardization. *Nec Technical Journal*, 3(3).
- Naylor, W. (2010). *Sharing Patient Health Information: A review of health information privacy and electronic health records in New Zealand*. Wellington: Cancer Control New Zealand .
- Nergiz, M. E., Atzori, M., & Clifton, C. W. (2007). Hiding the Presence of Individuals from Shared Databases. *International Conference on Management of Data* (pp. 665-676). Beijing, China: ACM SIGMOD.
- Nissenbaum, H. (2009). Privacy in Context: Technology, Policy, and the Integrity of Social Life. *Stanford University Press*.
- Nunamaker, J. F., Chen, J. M., & Purdin, T. D. (1991). Systems Development in Information Systems Research. *Journal of Management Information Systems* , 7(3), 89-106.
- Ogunyemi, O., Meeker, D., Kim, H., & Boxwala, A. (2013). Identifying Appropriate Reference Data Models for Comparative Effectiveness Research (CER) Studies Based on Data from Clinical Information Systems. *Medical Care*, 45-52.
- Olds, R. (2015). *The Virtual Health Information Network: options paper and implementation path*. New Zealand: The Health Information Network steering committee.
- Orencik, C., Selcuk, A., Savas, E., & Kantarcioglu, M. (2016). MultiKeyword search over encrypted data with scoring and search pattern obfuscation. *International Journal of Information Security*, 251–269.
- Osborn, S., Sandhu, R., & Munawer, Q. (2000). Configuring role-based access control to enforce mandatory and discretionary access control policies. *ACM Transactions on Information and System Security (TISSEC)*, 85-106.
- Oude, W., Velsen, L. v., Huygens, M., & Hermens, H. (2015). Requirements for and Barriers towards Interoperable eHealth Technology in Primary Care . *IEEE Internet Computing* , 10-19.
- Owopetu, O. O. (2013). *Private Cloud Implementation and Security*. Turku, Finland : Turku University of Applied Sciences .

- Patel, S., & Patel, A. (2016). A big Data Revolution in Health Care Sector: opportunities, Challenges, and Technological Advancements . *International Journal of Information Sciences and Techniques (IJIST)*, 155-162.
- Patton, M. (1990). Qualitative Evaluation and Research Methods. *Sage Publications*, 182-183.
- Patton, M. (2001). *Qualitative evaluation and research methods*. Newbury Park: Sage Publications.
- Patton, M. Q. (2002). *Qualitative research & evaluation methods (3rd Edition)*. CA: SAGE.
- Patton, M. Q. (2002). *Qualitative research and Evaluation: Chapter 7: Qualitative Interviewing (3rd Edition)*. Sage Publications.
- Pedersen, J. M., Riaz, M., Junior, J. C., Dubalski, B., Ledzinski, D., & Patel, A. (2011). Assessing Measurements of QoS for global Cloud Computing Services . *IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing* , 682-689.
- Peffers, K., Tuunamen, T., & Rothenberger, M. A. (2008). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), 45–77. doi:10.2753/MIS0742-1222240302
- Peixoto, H., Domingues, A., & Fernandes, B. (2016). Steps towards Interoperability in Healthcare Environment. In J. Machado, & A. Abelha, *Applying Business Intelligence to Clinical and Healthcare Organizations* (pp. 1-23). IGI Global .
- Pennic, F. (2015, February 02). *4 Challenges of Establishing EHR Interoperability*. Retrieved February 15, 2017, from HIT Consultant: <http://hitconsultant.net/2015/10/02/4-challenges-of-establishing-ehr-interoperability/>
- Phaphoom, N., Wang, X., Samuel, S., Helmer, S., & Abrahamsson, P. (2015). A survey study on major technical barriers affecting the decision to adopt cloud services. *The Journal of Systems and Software*, 167-181.
- Pramanick, N., & Ali, S. T. (2017). A comparative survey of searchable encryption schemes . *8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-15). Delhi, India : IEEE.

- Prasser, F., Kohlmayer, F., Lautenschläger, R., & Kuhn, K. A. (2014). ARX - A Comprehensive Tool for Anonymizing Biomedical Data . *Technische Universität München, München, Germany*, 984-993.
- Priyanga.P, & MuthuKumar.V.P. (2015). Cloud computing for healthcare organisation. *International Journal of Multidisciplinary Research and Development* , 487-493.
- Public Law . (1996). HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT OF 1996. *Public Law 104-191, 104th Congress*.
- Rao, R. V., & Selvamani, K. (2015). Data Security Challenges and Its Solutions in Cloud Computing. *International Conference on Intelligent Computing, Communication & Convergence* (pp. 204-209). Bhubaneswar, Odisha, India: Elsevier.
- Rashid, A. H., & Yasin, N. B. (2015, April). Privacy Preserving Data Publishing: Review. *International Journal of Physical Sciences*, 10(7), 239-247.
- Rashid, A. H., & Yasin, N. B. (2015, March). Sharing healthcare information based on privacy preservation. *Scientific Research and Essays*, 10(5), 184-195. doi:10.5897/SRE11.862
- Raval, D., & Jangale, S. (2016, September). Cloud based Information Security and Privacy in Healthcare. *International Journal of Computer Applications*, 150(4), 11-15.
- Rghioui, A., L'arje, A., Elouaai, F., & Bouhorma, M. (2014). The Internet of Things for Healthcare Monitoring: Security Review and Proposed Solution. *2014 Third IEEE International Colloquium in Information Science and Technology (CIST)*, (pp. 384 - 389).
- Rosado, D. G., Gómez, R., Mellado, D., & Fernández-Medina, E. (2012). Security Analysis in the Migration to Cloud Environments. *Future Internet*, 469-487.
- Rossmann, G. B., & Rallis, S. F. (2016). *An Introduction to Qualitative Research: Learning in the Field*. Sage.
- Rouse, M. (2015, June). *SearchHealthIT*. Retrieved February 25, 2017, from TechTarget: <http://searchhealthit.techtarget.com/definition/Clinical-Document-Architecture-CDA>
- Rouse, M. (2019). *Amazon Web Services (AWS)*. Retrieved from TechTarget: <https://searchaws.techtarget.com/definition/Amazon-Web-Services>
- Rouse, M. (2019). *cryptography*. Retrieved from TechTarget: <https://searchsecurity.techtarget.com/definition/cryptography>

- Rouse, Margaret. (2010, June). *Public Health Information Network*. Retrieved from SearchHealthIT: <http://searchhealthit.techtarget.com/definition/Public-Health-Information-Network>
- Rowley, J. (2002). Using Case Studies in Research. *Management Research News*, 25(1). Retrieved from http://www.psyking.net/HTMLobj-3843/using_case_study_in_research.pdf
- Sachdeva, S., & Bhalla, S. (2010). Semantic Interoperability in Healthcare Information for EHR Databases. *Graduate Department of Computer and Information Systems*, 157–173.
- Sahai, A., & Waters, B. (2005). Fuzzy Identity-Based Encryption. *Annu. Int. Conf. Theory Appl. Cryptograph. Techn.*, New York, USA: Springer.
- Samarati, P., & Sweeney, L. (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. SRI International.
- Sandhu, R. S., Coyne, E. J., Feinstein, H. L., & Youman, C. E. (1996). Role-based access control models. *IEEE Computer*, 38-47.
- Sangeetha, P., & Kavitha, M. (2016). Analysis of an effective, scalable and secured data sharing service in cloud computing. *International Journal of Modern Trends in Engineering and Research*, 135-141.
- Sargent, R. G. (2009). Verification and Validation of Simulation Models. *Winter Simulation Conference* (pp. 162-176). IEEE.
- Schneier, B. (2013). NSA Surveillance: a Guide to Staying Secure. *Schneier on Security*. Retrieved from: https://www.schneier.com/essays/archives/2013/09/nsa_surveillance_a_g.html
- Schneier, B. (2015). *Data Encryption Standard (DES)*. USA: John Wiley & Sons, Inc.
- Schutt, R. K. (2009). *Investigating the social world: The process and practice of research (6th Edition)*. CA: SAGE.
- SDMX-HD. (2016). Retrieved February 22, 2017, from Statistical Data and Metadata Exchange-Health Domain Standard Specification.
- Seidman, I. (2006). *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences (3rd Edition)*. New York: Teachers College Press.

- Shakir, M. (2002). The selection of case studies: Strategies and their applications to IS implementation cases studies. *Institute of Information and mathematical Sciences, Massey Univeristy, Albany* , 191-198.
- Shariati, S. M., Abouzarjomehri, & Ahmadzadegan, M. H. (2015). Challenges and security issues in cloud computing from two perspectives: Data security and privacy protection. *2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI)*. IEEE.
- Sharma, K., Jayashankar, A., Banu, K. S., & Tripathy, B. K. (2016). Data Anonymization Through Slicing Based on Graph-Based Vertical Partitioning. In K. Sharma, A. Jayashankar, K. S. Banu, & B. K. Tripathy, *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics* (Vol. 44, pp. 569-576). Springer India.
- Sinaci, A., & Erturkmen, G. B. (2013). A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. *Journal of Biomedical Informatics*, 784-794.
- Snowden, E. (2013, June). NSA whistleblower answers reader. Retrieved from <http://www.theguardian.com/world/2013/jun/17/edward-snowden-nsa-files-whistleblower>
- Song, D. X., Wagner, D., & Perrig, A. (2000). Practical techniques for searches on encrypted data. *Security and Privacy Proceedings* (pp. 44-55). IEEE.
- Spinola, M. (2009, September 6). The Five Characteristics of Cloud Computing. *Cloud Computing Journal: Cloud Expo Blog Feed Post*. Retrieved January 10, 2019, from <http://cloudcomputing.sys-con.com/node/1087426>
- Stake, R. (1995). *The art of case study research*. Thousand Oaks: Sage.
- Stake, R. E. (2013). *Multiple Case Study Analysis*.
- Stallings, W. (2005). *Cryptography and Network Security Principles and Practices, Fourth Edition*. Prentice Hall.
- Studnicki, J., Steverson, B., Myers, B., Hevner, A., & Berndt, D. (1997). Comprehensive assessment for tracking community health (CATCH). *Best Practices and Benchmarking in Healthcare : a Practical Journal for Clinical and Management Application*, 2(5), 196-207.
- Sultan, N. (2014). Making use of cloud computing for healthcare provision: Opportunities and challenges. *International Journal of Information Management*, 177–184.

Sung, M., & Pentland, A. (2004). Health and Lifestyle Networking through Distributed Mobile Devices. *WAMES 2004*, (pp. 15-17). Boston. Retrieved from http://lcawww.epfl.ch/luo/WAMES%202004_files/WAMESproceedings.pdf#page=15

Sweeney, L. (2002). k-Anonymity: A Model For Protecting Privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 557-570.

Taslakian, B., Sebaaly, M. G., & Al-Kutoubi, A. (2016). Patient Evaluation and Preparation in Vascular and Interventional Radiology: What Every Interventional Radiologist Should Know (Part 1: Patient Assessment and Laboratory Tests). *Cardiovasc Intervent Radiol* , 39(3), 325-333.

Tebaa, M., Hajji, S. E., & Ghazi, A. E. (2012). Homomorphic Encryption Applied to the Cloud Computing Security. *Proceedings of the World Congress on Engineering* , 1. London, U.K.

Tellis, W. M. (1997). Application of a Case Study Methodology. *The Qualitative Report*, 1-19.

Thomas, D. R. (2011). A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, 27(2), 237-246.

Thompson, T. G., & Brailer, D. J. (2004). The Decade of Health Information Technology: Delivering Consumer-centric and Information-rich Health Care: Framework for Strategic Action . *Department of Health & Human Services* .

Torabi, S., & Beznosov, K. (2013). Privacy Aspects of Health Related Information Sharing in Online Social Networks. *2013 USENIX Workshop on Health Information Technologies* .

Tsiknakis, M., Katehakis, D. G., & Orphanoudakis, S. C. (2002). An open, component-based information infrastructure for integrated health information networks. *International Journal of Medical Informatics*, 3-26.

Van-den Hoven, J., Blaauw, Martijn, Pieters, Wolter, Warnier, & Martijn. (2016). Privacy and Information Technology. In *The Stanford Encyclopedia of Philosophy (Spring 2016 Edition)* Edward N. Zalta (ed.).

Veer, H. v., & Wiles, A. (2008). Achieving Technical Interoperability - the ETSI Approach. *European Telecommunications Standards Institute*.

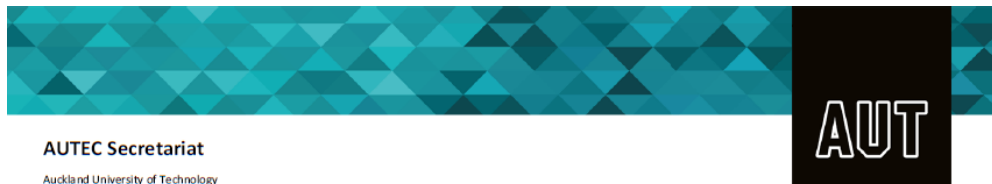
Vest, J. R., & Gamm, L. D. (2010). Health information exchange: persistent challenges and new strategies. *Journal of the American Medical Informatics Association*, 288-294.

- Victor, N., & Lopez, D. (2016). Privacy models for big data: a survey. *International Journal of Big Data Intelligence*, 61-75.
- Victor, N., & Lopez, D. (2016). Privacy models for big data: a survey. *Int. J. Big Data Intelligence*, 3(1), 61-75.
- Vikas, S., Gurudatt, K., Vishnu, M., & Prashant, K. (2013). Private Vs Public Cloud . *International Journal of Computer Science & Communication Networks*, 79-83.
- Vo, M.-T., Nghi, T. T., Tran, V.-S., Mai, L., & Le, C.-T. (2015). Wireless Sensor Network for Real Time Healthcare Monitoring: Network Design and Performance Evaluation Simulation. *5th International Conference on Biomedical Engineering in Vietnam*, 46, 87-91.
- Wager, K. A., Lee, F. W., & Glaser, J. P. (2013). *Health Care Information Systems: A Practical Approach for Health Care Management*, 3rd Edition. Jossey-Bass.
- Walker, J., Pan, E., Johnston, D., Adler-Milstein, J., Bates, D. W., & Middleton, B. (2004). The Value Of Health Care Information Exchange And Interoperability. *Center for Information Technology Leadership*, 1-176.
- Wang, B., Yu, S., Lou, W., & Hou, Y. T. (2014). Privacy-preserving multi-keyword fuzzy search over encrypted data in the cloud. *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications* (pp. 2112-2120). Toronto, ON, Canada: IEEE.
- Weber, R. H. (2015). Internet of things: Privacy issues revisited. *computer law & security review*, 31, 618–627.
- Weir, C. R., Hammond, K. W., Embi, P. J., Efthimiadis, E. N., Thielke, S. M., & Hedeem, A. N. (2011, August). An exploration of the impact of computerized patient documentation on clinical collaboration. *International Journal of Medical Informatics*, 80(8), 62-71. doi:10.1016/j.ijmedinf.2011.01.003
- West, M. (2015). A Comparative Analysis of HL7 and NIEM: Enabling Justice-Health Data Exchange. *Technical Brief*.
- Whiddett, R., Hunter, I., Engelbrecht, J., & Handy, J. (2006, July). Patients' attitudes towards sharing their health information. *International Journal of Medical Informatics*, 75(7), 530-541.
- Whitman, L. E., & Panetto, H. (2006). The missing link: Culture and language barriers to interoperability. *Annual Reviews in Control*, 30(2), 233–241.

- WHO. (2017). *Classifications*. Retrieved February 22, 2017, from World Health Organization: <http://www.who.int/classifications/icd/en/>
- WHO. (2017). *SNOMED CT to ICD-10 Cross-Map Technology Preview Release*. Retrieved February 22, 2017, from World Health Organization: <http://goo.gl/o0d8s>
- WHO. (2017). *Structure and principles*. Retrieved February 22, 2017, from Who Collaborating Centre for Drug Statistics Methodology: https://www.whocc.no/atc/structure_and_principles/
- Winans, T. B., & Brown, J. S. (2009). A collection of working papers. *Cloud Computing*, 2.
- Wong, R. C.-W., Li, J., Fu, A. W.-C., & Wang, K. (2006). (α , k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. *COMP 150 - Cryptography*, 754-759.
- World Health Organization . (2017, September). *European Health Information Initiative (EHII)*. Retrieved from World Health Organization, Retrieved from: <http://www.euro.who.int/en/data-and-evidence/european-health-information-initiative-ehii>
- Xia, W., Heatherly, R., Ding, X., Li, J., & Malin, B. (2015). R-U policy frontiers for health data de-identification. *J Am Med Inform Assoc*.
- Xiao, X., & Tao, Y. (2006). Personalized Privacy Preservation. *International conference on Management of data* (pp. 229-240). Chigago: ACM SIGMOD.
- Xu, Y., Ma, T., Tang, M., & Tian, W. (2014). A Survey of Privacy Preserving Data Publishing using Generalization and Suppression. *Applied Mathematics & Information Sciences*, 8(3), 1103-1116.
- Yang, G., Li, X., Mantysalo, M., Zhou, X., Pang, Z., Xu, L. D., . . . Zheng, L. (2014). Technologies and architectures of the Internet-of-Things (IoT). *IEEE Transactions on Industrial Informatics*, 10(4), 2180 - 2191.
- Yassein, M. B., Aljawarneh, S., Qawasmeh, E., Mardini, W., & Khamayseh, Y. (2017). Comprehensive study of symmetric key and asymmetric key encryption algorithms. *International Conference on Engineering and Technology (ICET)*. Antalya, Turkey: IEEE.
- Yimam, D., & Fernandez, E. B. (2016). A survey of compliance issues in cloud computing. *Journal of Internet Services and Applications*, 1-12.
- Yin, R. K. (1994). *Case study research: Design and methods (2nd ed.)*. Thousand Oaks: Sage.
- Yin, R. K. (2009). *Case Study Research: Design and Methods*. SAGE, 2009.

- Yu, S., Wang, C., Ren, K., & Lou, W. (2010). Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing. *INFOCOM, 2010 Proceedings IEEE*. IEEE.
- Yuan, E., & Tong, J. (2005). Attributed based access control (ABAC) for Web services. *IEEE International Conference on Web Services (ICWS'05)* (p. 569). Orlando, FL, USA: IEEE.
- Yüksel, B., Küpçü, A., & Özkasap, Ö. (2017). Research issues for privacy and security of electronic health services. *Future Generation Computer Systems*, 1-17.
- Zhang, R., & Liu, L. (2010). Security Models and Requirements for Healthcare Application Clouds. *IEEE 3rd International Conference on Cloud Computing (CLOUD)*,. IEEE.
- Zhang, Y., Chen, X., Li, J., Wong, D. S., Li, H., & You, I. (2017). Ensuring attribute privacy protection and fast decryption for outsourced data security in mobile cloud computing. *Information Sciences*, 42-6.
- Zhou, X., Ackerman, M. S., & Zheng, K. (2009). I just don't know why it's gone: maintaining informal information use in inpatient care. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York.

Appendix B: Ethic Application Approval from AUTECH



AUTECH Secretariat

Auckland University of Technology
D-88, WU406 Level 4 WU Building City Campus
T: +64 9 921 9999 ext. 8316
E: ethics@aut.ac.nz
www.aut.ac.nz/researchethics

31 July 2017

Jairo Gutierrez
Faculty of Design and Creative Technologies

Dear Jairo

Re Ethics Application: **17/228 Privacy-aware cloud-based architecture for sharing healthcare information**

Thank you for providing evidence as requested, which satisfies the points raised by the Auckland University of Technology Ethics Committee (AUTECH).

Your ethics application has been approved in stages for three years until 31 July 2020.

Standard Conditions of Approval

1. A progress report is due annually on the anniversary of the approval date, using form EA2, which is available online through <http://www.aut.ac.nz/researchethics>.
2. A final report is due at the expiration of the approval period, or, upon completion of project, using form EA3, which is available online through <http://www.aut.ac.nz/researchethics>.
3. Any amendments to the project must be approved by AUTECH prior to being implemented. Amendments can be requested using the EA2 form: <http://www.aut.ac.nz/researchethics>.
4. Any serious or unexpected adverse events must be reported to AUTECH Secretariat as a matter of priority.
5. Any unforeseen events that might affect continued ethical acceptability of the project should also be reported to the AUTECH Secretariat as a matter of priority.

Non-Standard Conditions of Approval

1. Amendment of the Information sheets as follows:
 - a. Include advice the interviews will be taped.
 - b. Expand the withdrawal statement (which can be found on the exemplar on the ethics website) to include withdrawal of data.

This approval is for the data collection stage of the research. Full information about future stages of this research needs to be provided to and approved by AUTECH before they commence.

Please quote the application number and title on all future correspondence related to this project.

AUTECH grants ethical approval only. If you require management approval for access for your research from another institution or organisation then you are responsible for obtaining it. You are reminded that it is your responsibility to ensure that the spelling and grammar of documents being provided to participants or external organisations is of a high standard.

For any enquiries, please contact ethics@aut.ac.nz

Yours sincerely,

Kate O'Connor
Executive Manager
Auckland University of Technology Ethics Committee

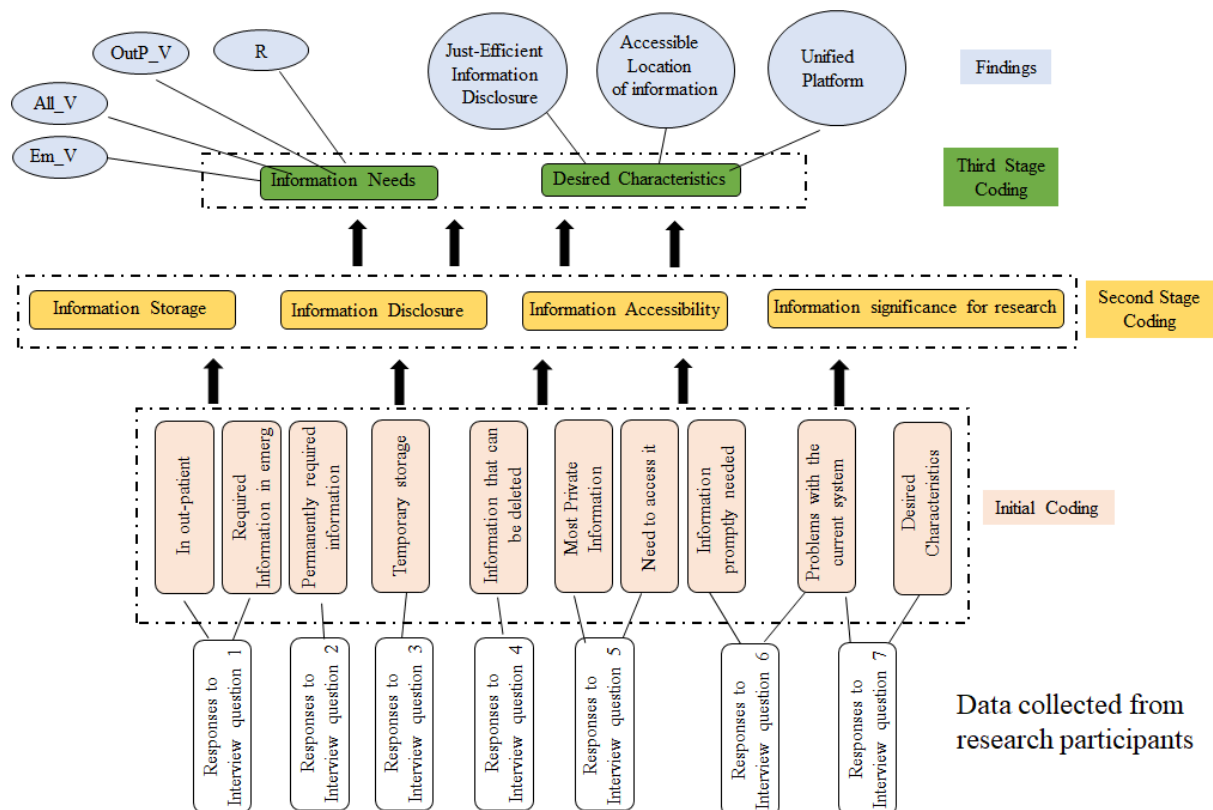
Cc: fa_d_jh@yahoo.com

Appendix C: Data Analysis and Coding

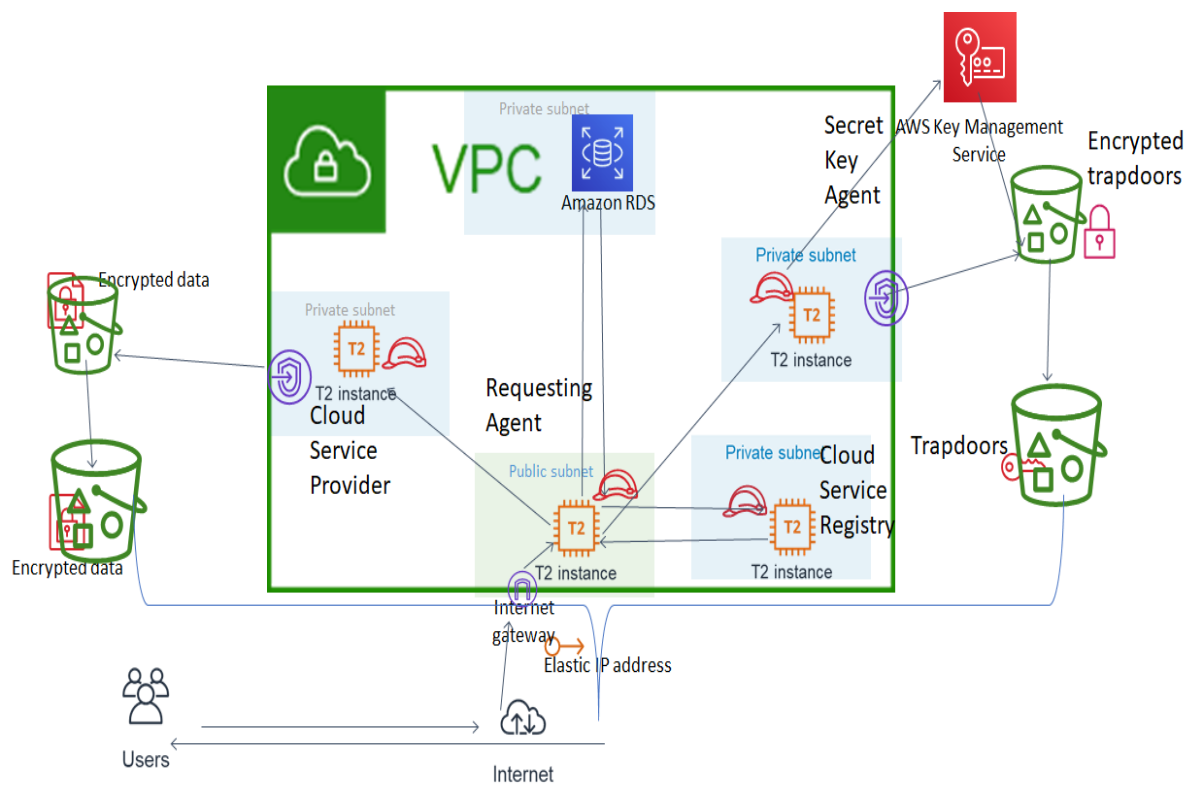
NVivo 12

My Data Analysis (NVivo 12).mp - NVivo 12 Plus

Name	Files	References	Created On	Created By	Modified On	Modified By
Question 1		16	16 19/03/2018 12:30 PM	FA	19/03/2018 12:30 PM	FA
In out-patient		4	4 10/04/2018 11:14 AM	FA	10/04/2018 11:43 AM	FA
Required information		16	95 10/04/2018 11:09 AM	FA	10/04/2018 11:45 AM	FA
Question 2		16	17 19/03/2018 12:30 PM	FA	10/04/2018 11:33 AM	FA
Permanently required information		16	64 21/04/2018 11:15 AM	FA	21/04/2018 11:45 AM	FA
Question 3		16	16 19/03/2018 12:30 PM	FA	19/03/2018 12:30 PM	FA
Not very important information		14	28 22/04/2018 12:15 PM	FA	22/04/2018 12:58 PM	FA
Question 4		16	16 19/03/2018 12:30 PM	FA	19/03/2018 12:30 PM	FA
Information that can be deleted		0	0 22/04/2018 1:00 PM	FA	22/04/2018 1:00 PM	FA
Question 5		16	16 19/03/2018 12:30 PM	FA	19/03/2018 12:30 PM	FA
Most Private Information		11	26 1/05/2018 10:07 AM	FA	1/05/2018 10:25 AM	FA
Need to access it		14	25 1/05/2018 10:08 AM	FA	1/05/2018 10:27 AM	FA
Question 6		16	16 19/03/2018 12:30 PM	FA	19/03/2018 12:30 PM	FA
Information promptly needed in emergenc		16	45 1/05/2018 10:53 AM	FA	1/05/2018 11:05 AM	FA
Problem with current system		1	1 1/05/2018 10:54 AM	FA	1/05/2018 10:54 AM	FA
Question 7		16	16 19/03/2018 12:30 PM	FA	19/03/2018 12:30 PM	FA
Characteristics		16	48 1/05/2018 11:06 AM	FA	1/05/2018 11:53 AM	FA
Problems with current system		9	15 1/05/2018 11:30 AM	FA	1/05/2018 11:53 AM	FA



Appendix D: AWS implementation diagram



Appendix E: Sample of the dummy dataset used

Number	ZipCode	Age	Gender	Nationality	MaritalStatus	BloodType	MedicalCondition	Severity	Treatment	Donor/Non-donor	Allergy
1	155214	89	Female	African	Married	B -	Heat Stress	3	CRG	Non-donor	none
2	155215	3	Male	Middle East	Single	O -	Abdominal Aortic Aneurysm	5	ERF	Donor	Aspirin allergy
3	155216	74	Female	African	Partnership	AB +	Sexually transmitted infection	3	BEU	Non-donor	Penicillin allergy
4	155217	98	Male	Black	Separated	O +	Hepatitis B	3	XRF	Donor	Latex allergy
5	155218	88	Female	White	Separated	B +	Back Belts	5	TCO	Non-donor	Ibuprofen allergy
6	155219	31	Female	American Indian	Widowed	O -	Birth Defect	2	GRT	Non-donor	Cat allergy
7	155220	26	Male	American	Partnership	O +	Cervical cancer	5	DCU	Donor	Peanut allergy
8	155221	49	Male	Russian	Partnership	B +	Hepatitis A	5	CCR	Donor	Cockroach allergy
9	155222	30	Female	Australian	Widowed	AB +	Blood cancer	3	CGD	Non-donor	Chemotherapy drugs allergy
10	155223	63	Male	Latino	Single	A -	Diabetes	1	TCO	Non-donor	none
11	155224	16	Male	Fiji	Single	B +	Oral cancer	4	VTT	Non-donor	none
12	155225	96	Male	Australian	Married	O -	Blood pressure	4	TTR	Non-donor	Peanut allergy
13	155226	16	Male	Australian	Single	AB +	Blood cancer	2	UV	Non-donor	Mold allergy
14	155227	7	Female	Pacific	Single	A -	Yellow fever	4	CGD	Donor	Mold allergy
15	155228	54	Male	New Zealander	Widowed	O +	Appendicitis	3	SSW	Non-donor	Skin allergy
16	155229	21	Female	Asian	Separated	AB +	Sexually transmitted infection	1	HUU	Non-donor	Mold allergy
17	155230	9	Female	White	Single	AB -	Brainerd Diarrhea	5	BTG	Non-donor	Dust allergy
18	155231	80	Male	Black	Widowed	O +	Liver cancer	3	GRT	Donor	Aspirin allergy
19	155232	5	Female	African	Single	A -	Hepatitis A	1	CQQ	Donor	Latex allergy
20	155233	5	Male	White	Single	A -	Hepatitis C	4	GFF	Non-donor	Mold allergy
21	155234	13	Male	Middle East	Single	B +	Hearing impairment	2	STT	Donor	Skin allergy
22	155235	52	Male	American Indian	Widowed	B -	Appendicitis	5	JED	Non-donor	none
23	155236	52	Female	Dutch	Separated	B -	None	2	EFT	Donor	Aspirin allergy
24	155237	68	Male	Indian	Widowed	O +	Hepatitis C	4	CT	Donor	Milk allergy
25	155238	6	Female	Scottish	Single	AB -	Diabetes	1	LUD	Non-donor	Cockroach allergy
26	155239	27	Male	New Zealander	Engaged	B +	Sexually transmitted infection	2	NHK	Donor	Ibuprofen allergy
27	155240	10	Male	Italian	Single	B -	Yellow fever	1	CWD	Non-donor	Rhinitis allergy
28	155241	31	Male	Indian	Engaged	B -	Birth Defect	3	EBB	Non-donor	Milk allergy
29	155242	27	Female	American	Partnership	O -	Back Belts	3	NHJ	Non-donor	Penicillin allergy
30	155243	68	Male	Asian	Partnership	AB +	Diabetes	5	CER	Donor	Mold allergy
31	155244	39	Female	Dutch	Partnership	B +	Birth Defect	4	XRF	Non-donor	Sulfa drugs allergy
32	155245	51	Male	Scottish	Single	AB -	Lung cancer	2	TRR	Donor	Penicillin allergy
33	155246	14	Female	Irish	Single	B +	Yellow fever	3	UUV	Non-donor	Mold allergy
34	155247	28	Male	New Zealander	Partnership	B -	Yellow fever	4	XIN	Donor	Sulfa drugs allergy
35	155248	51	Female	Middle East	Widowed	A -	Hepatitis C	3	XCQ	Non-donor	Milk allergy
36	155249	67	Male	Italian	Separated	O +	Abdominal Aortic Aneurysm	5	TRR	Donor	Skin allergy
37	155250	15	Female	Black	Single	O +	Malaria	2	XIN	Donor	Insect sting allergy
38	155251	67	Female	Russian	Married	AB +	Colon cancer	2	PIN	Non-donor	none
39	155252	55	Male	Filipino	Widowed	O -	None	3	NHK	Donor	Ibuprofen allergy
40	155253	13	Male	Indian	Single	A -	Sexually transmitted infection	5	TCS	Non-donor	Milk allergy
41	155254	10	Female	Irish	Single	A +	Back Belts	4	BGG	Non-donor	Chemotherapy drugs allergy
42	155255	84	Female	Latino	Single	AB -	Oral cancer	5	DCU	Donor	Eye allergy
43	155256	73	Male	Norwegian	Widowed	B +	Liver cancer	5	ERF	Non-donor	Milk allergy
44	155257	20	Male	Indian	Partnership	AB -	Birth Defect	5	IKH	Non-donor	none
45	155258	37	Female	Latino	Widowed	AB -	Adenovirus	1	EBB	Non-donor	Insect sting allergy
46	155259	70	Male	American	Single	O -	Heat Stress	3	VTT	Donor	Rhinitis allergy
47	155260	49	Female	New Zealander	Separated	A -	Cervical cancer	1	CT	Non-donor	Sulfa drugs allergy
48	155261	38	Female	Indian	Engaged	A +	Adenovirus	3	VEF	Donor	Chemotherapy drugs allergy
49	155262	92	Female	Asian	Widowed	O +	Adenovirus	1	CFF	Donor	none
50	155263	59	Male	Russian	Widowed	O +	Birth Defect	4	STT	Non-donor	Milk allergy
51	155264	81	Male	American	Widowed	AB -	Hepatitis B	4	CTA	Donor	Sulfa drugs allergy
52	155265	70	Male	Asian Indian	Married	A -	Hearing impairment	3	BGG	Non-donor	Insect sting allergy
53	155266	25	Female	Indian	Separated	A -	Adenovirus	1	NGF	Donor	Rhinitis allergy
54	155267	52	Female	German	Separated	O +	Oral cancer	2	VEF	Donor	none
55	155268	12	Male	Latino	Single	A -	Heart disease	5	UV	Non-donor	Insect sting allergy
56	155269	45	Female	Pacific	Married	A -	Malaria	1	BFF	Non-donor	Dust allergy
57	155270	25	Female	Mexican	Engaged	B +	Sexually transmitted infection	4	TRR	Donor	Ibuprofen allergy
58	155271	41	Female	Irish	Widowed	A +	Appendicitis	4	UIJ	Donor	Skin allergy
59	155272	53	Male	Norwegian	Married	B +	Adenovirus	3	KLJ	Donor	none
60	155273	93	Male	Filipino	Widowed	B +	Birth Defect	3	CT	Donor	none
61	155274	48	Male	Italian	Widowed	O +	Hepatitis C	1	CGD	Non-donor	Ibuprofen allergy
62	155275	41	Male	Dutch	Separated	A +	Lung cancer	1	CGD	Donor	Milk allergy
63	155276	85	Female	Russian	Married	AB -	Abdominal Aortic Aneurysm	4	NHJ	Donor	Insect sting allergy
64	155277	49	Female	New Zealander	Married	AB +	Rabies	4	JED	Donor	none
65	155278	91	Female	Middle East	Widowed	A +	None	1	NHK	Donor	none
66	155279	17	Female	German	Single	O +	Blood pressure	5	TUH	Donor	none
67	155280	17	Male	Russian	Single	AB -	Heat Stress	1	UV	Non-donor	none
68	155281	73	Female	Pacific	Single	AB -	Appendicitis	2	TRR	Non-donor	Dust allergy
69	155282	31	Female	German	Widowed	AB +	Hepatitis B	3	KCB	Donor	Mold allergy
70	155283	14	Male	Fiji	Single	B +	Hepatitis C	3	CWD	Non-donor	Penicillin allergy
71	155284	29	Male	American	Engaged	B +	Yellow fever	4	NHJ	Donor	Chemotherapy drugs allergy
72	155285	13	Female	Latino	Single	O -	Anxiety	3	RDE	Non-donor	Sulfa drugs allergy
73	155286	7	Male	White	Single	O -	Heart disease	2	CTV	Donor	Ibuprofen allergy
74	155287	20	Female	German	Engaged	AB -	Cervical cancer	1	CCR	Donor	Dust allergy
75	155288	59	Male	Asian	Married	B -	Alcohol addiction	1	BFF	Non-donor	Sulfa drugs allergy
76	155289	47	Male	Dutch	Married	A +	Hepatitis C	2	CT	Donor	Aspirin allergy
77	155290	53	Male	New Zealander	Married	O +	Hepatitis B	4	CNY	Non-donor	Dust allergy
78	155291	82	Male	American Indian	Separated	A +	None	1	CCR	Non-donor	none
79	155292	40	Female	Pacific	Widowed	AB +	Colon cancer	2	TCO	Non-donor	Dust allergy
80	155293	30	Female	Indian	Widowed	A +	Abdominal Aortic Aneurysm	5	CCR	Non-donor	Aspirin allergy
81	155294	34	Female	American	Married	A +	Hepatitis B	2	CTA	Non-donor	none
82	155295	6	Female	Pacific	Single	A +	Alcohol addiction	4	CWF	Donor	Dust allergy
83	155296	81	Female	Middle East	Widowed	A -	Birth Defect	3	CCR	Non-donor	Eye allergy
84	155297	97	Female	Black	Separated	O -	Anxiety	2	CTV	Non-donor	Eye allergy
85	155298	64	Male	Fiji	Widowed	AB -	Blood cancer	3	UUV	Non-donor	Eye allergy
86	155299	100	Male	Norwegian	Single	A -	Malaria	4	CWF	Donor	Cat allergy
87	155300	27	Male	Australian	Separated	O +	Blood pressure	1	ERF	Donor	Cockroach allergy
88	155301	100	Female	Irish	Single	B +	Yellow fever	2	VEF	Donor	Peanut allergy
89	155302	22	Female	White	Married	B +	Birth Defect	2	UUV	Donor	Insect sting allergy
90	155303	56	Male	Black	Widowed	B -	HIV	3	BTG	Donor	Milk allergy
91	155304	56	Male	Pacific	Partnership	B -	Birth Defect	2	TCS	Donor	Insect sting allergy
92	155305	19	Female	Asian Indian	Married	B -	Oral cancer	1	CFF	Non-donor	Aspirin allergy
93	155306	31	Female	Irish	Married	A +	Alcohol addiction	5	XEE	Donor	Dust allergy
94	155307	13	Male	American Indian	Single	B -	Swollen glands	4	JED	Non-donor	Peanut allergy
95	155308	55	Female	Latino	Widowed	O +	Abdominal Aortic Aneurysm	3	TCS	Donor	Cockroach allergy

Appendix F: ARX Anonymized dataset sample

	ZipCode	Age	Gender	BloodType	MedicalCondition	Severity	Treatment	Target	Cockroach
1	125033	59	Female	AB -	Blood cancer	5	NHU	Target	Cockroach i
2	124633	54	Female	O -	Cervical cancer	2	HUU	Target	Skin allergy
3	124699	30	Female	A +	Cervical cancer	3	RH	Target	Cockroach i
4	124671	56	Female	O +	Cervical cancer	5	NGF	Target	Penicillin all
5	124779	56	Female	B +	Cervical cancer	3	ZEG	Target	Chemother
6	124577	58	Female	A +	Cervical cancer	3	ZEG	Target	Aspirin allen
7	124669	32	Female	AB -	Colon cancer	1	E88	Target	none
8	124515	53	Female	B -	Colon cancer	3	KLJ	Target	Penicillin all
9	124955	56	Female	O +	Diabetes	5	GFF	Target	Rhinitis aller
10	124522	59	Female	O +	none	0	none	Non-Target	none
11	124524	55	Female	B +	none	0	none	Non-Target	Skin allergy
12	124562	57	Female	A -	none	0	none	Non-Target	Rhinitis alle
13	124568	50	Female	B +	none	0	none	Non-Target	Penicillin all
14	124608	56	Female	B -	none	0	none	Non-Target	Sulfa drugs
15	124907	53	Female	B -	none	0	none	Non-Target	Chemother
16	125126	55	Female	A -	none	0	none	Non-Target	Cockroach i
17	125134	58	Female	AB +	none	0	none	Non-Target	Late allergy
18	125235	51	Female	AB -	none	0	none	Non-Target	Dust allergy
19	125266	56	Female	A -	none	0	none	Non-Target	Late allergy
20	125316	57	Female	A +	none	0	none	Non-Target	none
21	125331	58	Female	A +	none	0	none	Non-Target	Insect sting
22	125371	59	Female	A -	none	0	none	Non-Target	none
23	125428	59	Female	O +	none	0	none	Non-Target	none
24	124653	86	Female	O +	Blood cancer	2	HUU	Target	Cat allergy
25	125131	85	Female	B -	Blood cancer	5	UJI	Target	Late allergy
26	125155	87	Female	A -	Blood cancer	1	XJN	Target	none
27	124823	83	Female	A -	Blood pressure	4	VIT	Target	Milk allergy
28	125330	88	Female	B -	Blood pressure	4	BEU	Target	none
29	125155	84	Female	B +	Blood pressure	1	KLJ	Target	Skin allergy
30	124520	85	Female	O +	Colon cancer	1	XJN	Target	Ibuprofen al
31	125042	82	Female	AB +	Diabetes	2	TCS	Target	none

Results of ARX that indicate the possibilities of performing successful privacy attacks on datasets anonymized using the proposed model.

