A NoSQL-Based Framework for Managing and Integrating Genetic Disorders Data from Genetic Clinics and Research Centres in Saudi Arabia

Halima Edris Samra
Master of Computer Science

A thesis submitted in total fulfilment of the requirements for the degree of Doctor of Philosophy

College of Science, Health and Engineering
School of Engineering and Mathematical Sciences
Department of Computer Science and Information Technology
La Trobe University
Victoria, Australia

October 2020

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms

ICT	Information and Communication Technologies
Backend	Server-side: database management system
BSON	Binary JSON
DBLC	Database Lifecycle
DBMS	Database Management System
DMS	Data Management System
EHR	Electronic Health Record
EMR	Electronic Medical Record
ERD	Entity Relational Diagram
ETL	Extraction, Transformation and Load
Frontend	Client-side: Web interface application
G3DMS	Genetic Disorders Diagnosis Data Management System
GENE2D	An integrated data repository of genetic disorders data
HIPAA	Health Insurance Portability and Accountability Act
HIS	Health Information System
HIT	Health Information Technology
JSON	JavaScript Object Notation
MOH	Ministry of Health
NoSQL	Not Only SQL
OLAP	Online Analytical Processing
PACER-HD	Princess Al-Jawhara Centre of Excellence in Research of Hereditary Disorders
SDLC	Systems Development Lifecycle
SQL	Structured Query Language

Abstract

Current health information systems used in genetic research centres and clinics in the Kingdom of Saudi Arabia do not allow researchers and healthcare physicians to use genetic and clinical data in their research. There are also a few sources of clinical and genetic data for use in research in Saudi Arabia. Difficulties in implementing health information systems in the Saudi health system affect healthcare professionals in their daily workflow as well as their research contribution due to a lack of quality data. Data collection, data storage and a lack of system interoperability are major impediments facing healthcare professionals in managing the data required for their research. In addition, there are difficulties in integrating data from silos and scattered sources to provide standardised access to large datasets for patients with common health conditions.

The provision of genetic data for clinical research is a significant area of study. To date, there has been little investigation into a foundation for national genetic disorders databases. There is a need for a health informatics framework that facilitates diagnostic workflow and aids in decision making, enabling information reuse in medical research and public health. To address this need, the thesis first investigated the challenges facing Saudi physicians in clinical research, such as data provision, integration and sharing among Saudi hospitals. The study findings highlighted critical obstacles and revealed the current process of conducting clinical research in such a difficult environment. Based on these findings, this thesis aims to develop an efficient system to contribute to the process of managing data while diagnosing genetic conditions in Saudi genetic clinics and research centres. The novel genetic disorders diagnosis data management system called G3DMS was designed to be applicable in any genetic clinic or research centre. A further aim was to solve the issue of data integration to provide large datasets for clinical research from multiple sources of genetic diagnostic data systems. An integration framework was incorporated into the G3DMS using an integrated data repository based on a NoSQL database which generates an integrated data repository of genetic disorders data called GENE2D.

Statement of Authorship

Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis submitted for the award of any other degree or diploma. No other person's work has been used without due acknowledgement in the main text of the thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution.

Halima Samra
30 October 2020

Acknowledgments

All praise and glory be to Allah the Almighty God, the most gracious, the most merciful, for giving me the strength, opportunity and determination to accomplish my dreams and be able to complete the journey of PhD study. My grateful appreciation goes to my parents, husband, children and family. I wish to express my deepest gratitude to the King Abdulaziz University, Kingdom of Saudi Arabia, for awarding me a scholarship and for providing financial support throughout my studies.

I would like to express my grateful appreciation to my advisor, Professor Ben Soh, for his outstanding guidance, support and encouragement during the research work. His experience, immense knowledge and patience were supportive when facing difficulties during the PhD journey. His insightful and invaluable contributions were the driver that brought this thesis to its current form and content. My sincere thanks go to my co-supervisor, Dr Alice Li, for her support during my PhD study.

I would like to express my thanks and appreciation to Professor Jumana Al-Aama, Chair of the Department of Genetic Medicine and Director of the Princess Al-Jawhara Al Brahim Centre of Excellence in Research of Hereditary Disorders at King Abdulaziz University, Jeddah, Saudi Arabia. Her contribution and cooperation laid the foundation for this work.

I give special thanks to Dr Deema Jiaid and to the PACER-HD staff who participated in the focus group and during system testing for their valuable information, feedback and suggestions.

I would like to express my sincere appreciation to Dr Murad Karrar for his consultation and professional advice, to Dr Ahlam Mazi for her feedback on the questionnaire; and to Dr Naseem Alyahyawi, Dr Muna Alyahyawi and Dr Huda Alyahyawi for their support in distributing the survey to staff.

I would like to extend a special thanks to Dr Nora Al-Ghamdi for her assistance in the early stages of my PhD research.

Dedication

This work is dedicated to my parents, who gave me the opportunities and experience that have made me who I am. To my father, Edris Samra, my honourable teacher, and my role model in teaching and my constant motivation to complete my higher education, thank you for your inspiration. To my mother, I recognise her patience at me being away from her during my study, and for her sincere prayers and words of motivation.

I dedicate this thesis with love and gratitude to my husband and my partner in this achievement, Eng. Saeed Alyehyawi, for his unbounded support, continued motivation and presence by my side during the ups and downs of this journey. With his affection, encouragement, understanding and patience, I was able to complete my thesis.

To my heroes, my beloved children, Tahir, Tamir, Rafal and Bassam, each of you faced his own educational journey with courage and determination yet created the right atmosphere for me to pursue this research. I am so proud of each of you for your achievements; we all shared the journey and will share the rewards. Congratulations Tahir and Tamir on your Masters of Business ERP, Bassam on your Victorian Certificate of Education (VCE) achievements, and soon Rafal on your Bachelor of Biomedical Science (laboratory medicine). I express my deep and sincere gratitude to you for being such great companions on this journey.

To my wonderful gentlemen, my beloved brothers Abdul, Othman and Yasser, I am grateful to you all for always being there for me. I would like to express my gratitude and appreciation to you and your families for the continuous encouragement and unconditional support to my family and to me. You have done your best to bring an entertaining aspect to our journey away from the atmosphere of study, well done.

To my family-in-law in Jeddah, my father-in-law Mr. Yahya and auntie Hawa, warm thanks and appreciation for your sincere prayers and words of motivation.

My dedication extends to my beloved aunties, and all of my extended family in Australia; thank you for your encouragement and emotional support.

List of Publications

Published papers

1. **Samra**, H.; Li, A.; Soh, B.; Al Zain, M. Utilisation of hospital information systems for medical research in Saudi Arabia: A mixed-method exploration of the views of healthcare and IT professionals involved in hospital database management systems. Health Inf. Manag. J., **2019**, 49, 1–10. [Journal Impact Factor,1.833]

   This paper forms the basis of Chapter 3 and has been updated and rewritten with new content and structured to form a thesis chapter.

2. **Samra**, H.; Li, A.; Soh, B. G3DMS: Design and implementation of a data management system for the diagnosis of genetic disorders. Healthcare, **2020**, 8, 196. [Journal Impact Factor,1.916]

   This paper forms the basis of Chapter 7 and Chapter 8 and has been updated and rewritten with a thesis structure and organisation.

3. **Samra**, H.; Li, A.; Soh, B. GENE2D: A NoSQL integrated data repository of genetic disorders data. Healthcare, **2020**, 8, 257. [Journal Impact Factor,1.916]

   This paper forms the basis of Chapter 9 and has been updated and rewritten with a thesis structure and organisation.

4. **Samra**, H.; Li, A.; Soh, B. Design of a Clinical Database to Support Research Purposes: Challenges and Solutions. International Journal of Advanced and Applied Sciences, 8(3) 2021, Pages: 21-29.

   This paper forms the basis for the design of clinical databases in Chapter 5 and has been updated and rewritten with the thesis structure and organisation.

Accepted papers

Samra, H.; Li, A.; Soh, B.; Al Zain, M. Review of Contemporary Database Design and Implication for Big Data. International Journal of Smart Education and Urban Society (IJSEUS) (**Accepted**, October 2020).

This paper forms the basis for the general design of databases in the big data environment in Chapter 5 and has been updated and rewritten with the thesis structure and organisation.

Chapter 1: Introduction

1.0 Chapter overview

This chapter is an introduction to the thesis, providing an overview of the thesis content and organisation, and the relationship between chapters. Section 1.1 presents a general introduction to health information systems applications in healthcare, Section 1.2 demonstrates health information systems background challenges in medical research, and Section 1.3 provides an overview of health information systems challenges in Saudi Arabia. Next, Section 1.4 outlines the research motivation, and Section 1.5 discusses the research objectives, aims, significance, research approach and methodology. Section 1.6 presents the thesis contributions to knowledge and the research community. Finally, Section 1.7 outlines the thesis structure, and Section 1.8 concludes the chapter.

1.1 Introduction

This thesis focuses on the use of healthcare data to improve the care process, assist clinical decisions and enhance research studies. The research study was based on concepts and methods of computer science and information for developing frameworks for the use of information in healthcare operations and research as well as the creation of an integrated data repository.

Advances in health information technology and analytics increase the potential utility of clinical data to bridge the gap between research and practice in the medical field. Health information systems, such as the electronic medical record (EMR) and the electronic health record (EHR), are used to manage and analyse patient data throughout the care process in different healthcare environments. Vast amounts of data electronically stored during patient care have great potential for retrospective observational studies [1]. However, the way the data are captured and managed by various health information systems results in data fragmentation, and information is siloed in different databases and repositories [2]. Data quality issues in these systems make data analysis and interpretation a challenging method for a patient or population-level research [3]. In addition, technical, cultural and legal barriers hinder data integration and sharing, and there are privacy concerns [2]. The problem of clinical data provision for researchers in medical studies in Saudi Arabia exists at multiple levels. First, there is inadequate implementation of electronic medical records, a lack of a standard format of data collection, and poor traditional data collection and management methods used for medical research [4]. Second, there is a lack of standards and integration profiles to enable information exchange among healthcare information systems, as the healthcare system in Saudi Arabia is delivered by multiple agencies, both governmental and private organisations, with different data protection and sharing policies [5], [6]. This has prevented the establishment of a national database for genetic disorders as well as a database for genetic variations reported by diagnostic and research laboratories [7].

This thesis investigates the Saudi context to identify the major issues that hinder the exploitation of healthcare data sources such as electronic medical records and electronic health records for medical research. Accordingly, it proposes a solution for data management based on health informatics framework supports (clinical workflow, clinical decision, and reuse data for research); and a data integration framework based on a NoSQL document database to create an integrated data repository from multiple clinical sources.

1.2 Background

A health information system is used in healthcare to facilitate the clinical care process by capturing, storing, processing and delivering information to decision-makers. It is an important component of hospital information system solutions, which include the electronic medical records system, computerised physician order entry system, laboratory information system, pharmacy information system and radiology information system [8], as well as the emerging electronic health records in care settings which comprise a digitised version of all clinical data relevant to patients' care such as demographics, medical history, medication, care plan, laboratory data, radiology reports, physicians' feedback and billing information [3]. The proper application of health information systems has the potential to improve healthcare delivery and enable decision-makers to take informed action based on integrated data. The successful development and implementation of health information systems can improve the efficiency and effectiveness of healthcare services and outcomes [9]. Advances in health information technology and analytics increase the potential utility of clinical data to bridge the gap between research and practice in the medical field. Vast amounts of data electronically stored during patient care have great potential for retrospective observational studies [1]. However, barriers to the implementation of health information systems result in variation in adoption among organisations or even countries around the world [10]. The adoption rate is affected by different factors that influence the implementation outcome of health information systems, which can be categorised as ethical, financial, functionality, organisational, political, technical and training [11].

Medical and clinical research relies mostly on data collected specifically for the purpose of research or from the health information system through its multiple channels in healthcare organisation research [2]. There are multiple data collection approaches used to extract data from multiple sources for research. Healthcare data can be obtained using primary data collection methods such as observations, surveys and interviews as well as secondary sources of data which are pre-organised healthcare data from the electronic patient record. The secondary data collection methods can provide unlimited data but with many concerns about data reliability and usefulness [12]. The quality of data collected through health information systems such as electronic health records has been questioned due to incomplete, inconsistent and noisy data [13]. The collected data can be difficult to manage for research purposes unless it undergoes a preparation process to make it useful for research [14]. In general, the available data collection methods for research can be either labour intensive and time-consuming such as manual data collection from paper-based sources, or result in data quality issues such as direct extraction from health information systems [15].

In general, health information systems in any health organisation comprise different software solutions for data collection, storage and management. Data reside in different locations within the same organisation in various formats and structures, making data manipulation a complicated task. Lack of medical records integration among an organisation's specialisations or across institutions creates great difficulties in data analysis and medical research studies [16]. In addition, the current patient clinical care databases are often inadequate to assess health interventions because data are often missing or incorrect, and it is difficult to link patients' demographics and clinical care procedures [17]. Therefore, healthcare institutions need fundamental changes in the infrastructure and mechanisms for data collection, storage and exchange. For data to be suitable for use in research, systems must have the ability to be integrated to create a comprehensive view and provide meaningful insights for patients and researchers. But it is very challenging to integrate databases with different specifications on data models, database schemas, the queries they support, and the terminologies they use. Data sharing through databases is more common practice for clinical research as data collected at multiple sites are integrated with some disease-oriented database systems, since one location may not be able to collect sufficient data for analysis. Clinical institutions may also be limited in terms of research interests so that a common database can make the collected data available to researchers in a variety of locations [14]. The process of data integration requires bringing together scientific methods and specifications needing to be stored in a database. The overall goal of data integration for the clinical research community is to be able to answer questions about aggregated data which can be very difficult if each individual data source must be accessed separately or sequentially [18].

1.3 Overview of the Saudi context

Saudi Arabia is a developing country that, over the past few years, has achieved significant progress in broadening the adoption of health information technology. The application of health information systems in Saudi hospitals varies, as some hospitals lack computerised systems, and other hospitals are using systems from different vendors [19]. For example, the adoption rate of electronic health and medical records in Saudi Arabia varies according to the hospital size, type and ownership across all regions; however, multiple challenges delay full implementation. Despite the effort from the government to encourage health organisations to accelerate the adoption process of health information systems, barriers arise in system implementation [20]. Technical, administrative, financial and human factors are identified as critical barriers to successful implementation [21]. Difficulties in health information system implementation result in many issues in the Saudi health system and affect healthcare professionals in their daily workflow as well as their research contribution due to a lack of quality data. Data collection, data storage and a lack of system interoperability are the major impediments facing healthcare professionals in managing the data required for their research [22]. In addition, poorly functioning health information systems and lacking inter-integration were a major obstacle to developing centralised national databases for research use [7].

1.4 Research motivation

The provision of genetic data for clinical research is a significant area of study. To date, there has been little investigation into a foundation for Saudi genetic disorders databases. Also, there is a need for a health informatics framework that facilitates diagnostic workflow and aids decision making, enabling information reuse in medical research and public health. Motivated by the initiative to contribute to the field of health information management and integration, the thesis focuses on the area of genetic data provision for medical research in Saudi Arabia for the following reasons:

Saudi Arabia has one of the highest consanguinity rates in the world as first cousin marriages constitute 60–70% of all marriages [23]. This has increased the prevalence of rare inherited genetic disorders in Saudi Arabia, which is nearly double the overall rate in Europe and the United States, and ten times higher for specific diseases.

Although the government supports many projects and programs that aim to collect genetic data such as the Saudi Human Genome Program to help research and develop preventive measures to limit the prevalence of diseases in the region [24], the lack of uniformity and data silos due to individual efforts to obtain clinical and genetic data through centres and clinics has made it more difficult for researchers to pre-process and manage data associated with their studies [6].

Genetic clinics and research centres have helped to reveal genetic disorders and generate curated data which can be a valuable source for researchers if the data are collected properly and stored efficiently. Sharing such data could advance research and enhance healthcare quality by providing accurate information on commonly encountered inherited disorders and lower the incidence of genetic diseases using preventive measures such as premarital and preimplantation testing [6]. However, these centres lack adequate systems for managing and archiving patient data for research, making it difficult to integrate and build a national genetic diseases database.

Internal problems within Saudi clinics and genetic research centres in terms of data collection and retention for use in research hinder data integration and sharing for research and the establishment of a unified database of genetic diseases [7].

1.5 Research objectives and aims

The purpose of this study is to support genetic disorders diagnosis and research processes within an individual institution and allow clinical and genetic data integration among these clinics and centres to establish an integrated data repository of data on genetic disorders of Saudi patients. The thesis has the following objectives:

Review the literature about the current state of e-health and health informatics in Saudi Arabia on the use of health information systems in delivering care, decision making and reuse in research. The aim is to define the problem domain and identify the existing gaps in the research in this area.

Conduct a pilot study to help formulate the problem and identify possible challenges Saudi physicians may face in the use of health information systems in medical research. This study will help narrow the thesis focus on a significant area of great concern in the medical field in Saudi Arabia, the diagnosis of genetic conditions.

Design and implement a data management system for a genetic clinic and research centre considering a health informatics framework. The aim of the system is to support the diagnosis workflow, assist physicians in making informed diagnosis decisions based on the available data, and allow secondary use of data for research.

Design an integrated genetic disorders data repository extracted from multiple genetic clinics and research centres using efficient, cost-effective technologies. The aim is to support genetic studies and public health research in Saudi Arabia as well as establish the foundation for a Saudi specific genetic disorders database.

1.5.1 Research significance and implications

The thesis significance lies in the main objectives of this research to design a novel data management system for the diagnosis of genetic disorders with promising hybrid models which combine the benefits of capturing data as part of routine care, facilitating diagnosis decisions, and increasing the efficiency of research using an existing database. This system will assist physicians and researchers in recording and discovering rare and prevalent genetic cases in the Saudi population, resulting in a greater contribution to gene discovery globally and increasing Saudi submissions to the public archive of human genetic variants such as the ClinVar database. In addition, this thesis provides a significant contribution to the literature in Saudi Arabia and adds to the knowledge of the current challenges on the use of health information systems for medical research, and specifically, the obstacles Saudi researchers face in performing medical research. This research also discusses the design methods of clinical databases from previous studies and critically evaluates the traditional methods and their applicability to serve research. It investigates integration approaches thoroughly and evaluates each approach for their suitability for a low-resource setting and limited information and communication technology infrastructure capabilities. This work presents effective methods for designing a data management system and integration framework suitable for healthcare institutions with limited infrastructure and financial capabilities such as in developing countries. This research uses promising technologies such as a NoSQL database for integration as well as health informatics frameworks for evaluation. The results of this thesis can be implemented in any genetic clinic or research centre in Saudi Arabia and are cost-effective and easy to implement in developing countries.

1.5.2 Problem solving approach using research questions

This thesis follows the approach of the software development process through the three stages of development: analysis, design and implementation [25]. The three developmental stages are used to achieve the research aims and answer the research questions. Research questions are structured in an appropriate manner according to their relevance in each stage. Figure 1.1 shows the three stages as presented by Zelkowitz et al. and the research questions [25].

I. The problem analysis stage →

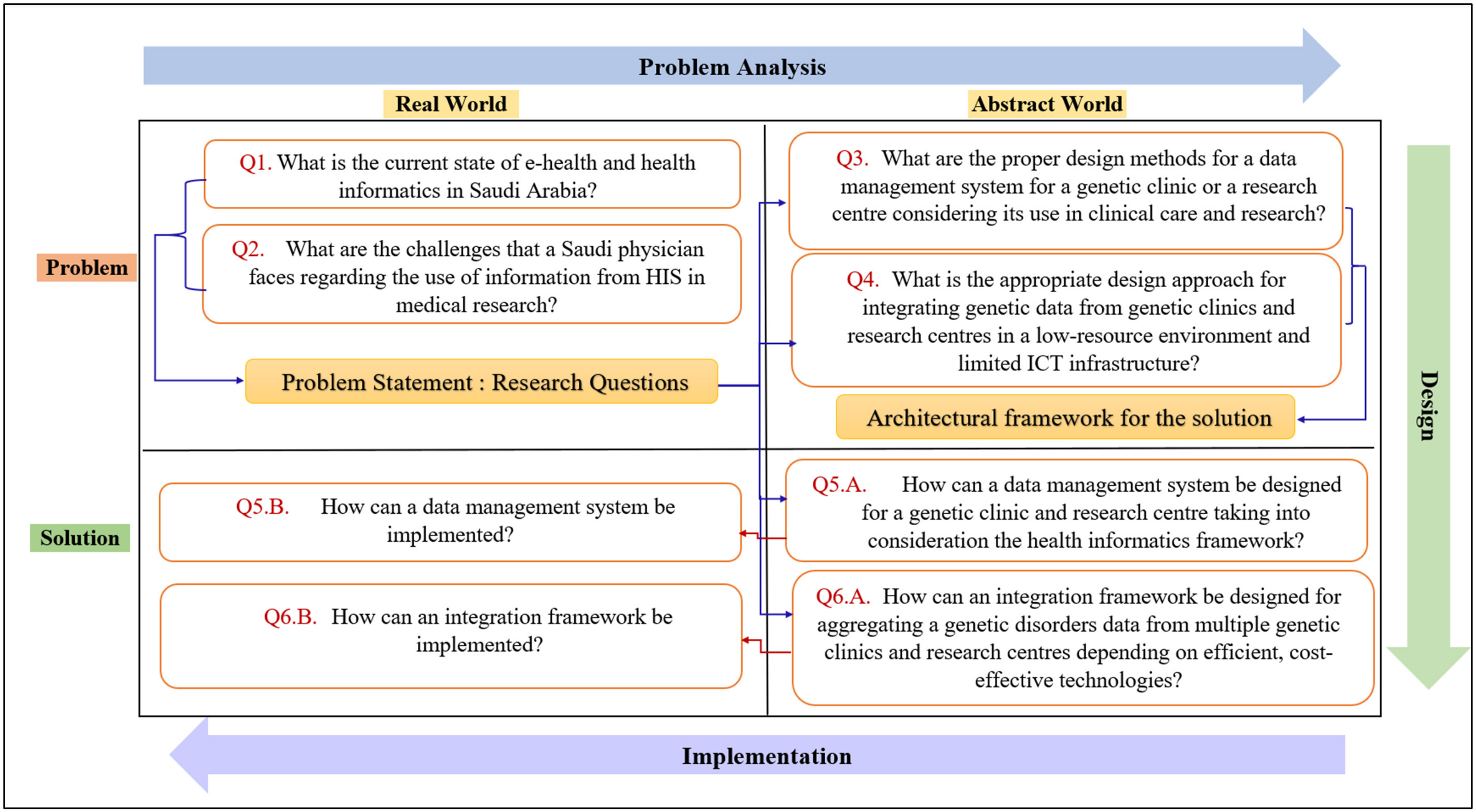
The aim of this stage is to define and understand the problem domain. Therefore, question 1 covers all aspects of the status of health information technologies in Saudi Arabia. Then, question 2 is used to narrow the scope of the research and focus on the use of health information systems in medical research. The study focuses on a significant area of genetic conditions. Next, the identified gaps in the literature and the findings from the pilot study help to define the problem statement. The problem statement poses four fundamental questions to cover the What and How aspects of the theoretical design concepts and physical models. Questions 3 and 4 are designed to help establish a thorough investigation and discuss the theoretical foundation behind each solution concept. Critical evaluation of possible solution approaches is based on the influencing factors from the implementation environment.

II. The design stage ↓

The aim of this stage is to abstract the problem and find a solution. This stage overlaps with the analysis stage; therefore, questions 3 and 4 falls in both stages. At this point, the solution framework is decided, and the systems architecture is defined. Next, question 5 on the design and implementation of the data management system is divided into question 5.A to guide the design lifecycle processes of the data management system through its early phases and question 5.B to proceed with the development in the latest phases. Also, question 6 on the design and implementation of the integration framework is divided into 6.A to lead the design steps of the integration framework and question 6.B to focus on the deployment phase.

III. The implementation stage ←

The aim of this stage is to follow the product from design in the abstract world to application in the real world. This stage also overlaps with the design stage; therefore, questions 5 and 6 are the guide throughout this stage, starting from modelling and coding to deployment.


Problem Analysis
Real World
Abstract World
Problem
Solution
Design
Implementation
Problem Statement : Research Questions
Architectural framework for the solution

Q1. What is the current state of e-health and health informatics in Saudi Arabia?

Q2. What are the challenges that a Saudi physician faces regarding the use of information from HIS in medical research?

Q3. What are the proper design methods for a data management system for a genetic clinic or a research centre considering its use in clinical care and research?

Q4. What is the appropriate design approach for integrating genetic data from genetic clinics and research centres in a low-resource environment and limited ICT infrastructure?

Q5.A. How can a data management system be designed for a genetic clinic and research centre taking into consideration the health informatics framework?

Q5.B. How can a data management system be implemented?

Q6.A. How can an integration framework be designed for aggregating a genetic disorders data from multiple genetic clinics and research centres depending on efficient, cost-effective technologies?

Q6.B. How can an integration framework be implemented?

Figure 1.1: Problem solving approach using research questions

1.5.3 Overall research methodology

The thesis methodology also uses the approach of the software development process [25] as presented in Section 1.5.2. The research methodology describes the answers in term of methods and approaches used as a response to the previous research questions presented in Figure 1.1. Figure 1.2 demonstrates the thesis methodology that describes each development stage, its associated research tasks, and the relevant chapters in the thesis structure.

I. The problem analysis stage ⇨

In this stage, the aim is to identify the problem in the real world by conducting a comprehensive investigation using a literature review in Chapter 2 to define the problem domain. Mixed methods (questionnaire, interviews and expert opinion) are used to perform a pilot study in Chapter 3 to understand the nature of the problem in medical research and narrow the scope of the thesis. The findings from the pilot study and the gaps identified in previous studies are used to define the problem statement, and research questions derived from the problem statement are discussed using a variety of techniques (analysis, comparisons and evaluation) to identify possible solution approaches and suggest the best approach for each design and then provide a solution framework in Chapter 4. A comprehensive review of theoretical background information about the proposed system design is presented in Chapter 5. The outcome of this stage is the solution architectural framework which is illustrated in Chapter 6.

II. The design stage ⇩

This stage involves converting the proposed theoretical model into a realistic design using theories and the methods chosen. The design stage overlaps the analysis stage; therefore, Chapter 4 and Chapter 6 are the foundation for the following design and implementation chapters of the thesis. The Barker method is used for the full system design lifecycle from design to implementation, and qualitative evaluation methods of the multilevel service design method and the informatics evaluation framework are used to evaluate the system design lifecycle process of the G3DMS. The first three phases of the Barker system design lifecycle of the genetic disorders diagnosis data management system called G3DMS are presented in Chapter 7. The integration framework follows the physical integration approach for the design of the proposed integrated data repository of genetic disorders data called GENE2D, and the NoSQL document database, MongoDB, is used as a backend. The GENE2D complete design and implementation processes are presented in Chapter 9.

III. The implementation stage

The design and the implementation stages overlap as well, starting from the design lifecycle of the development of G3DMS with design in Chapter 7 and implementation and evaluation in Chapter 8 to the integration framework of G3DMS and the design and implementation of GENE2D in Chapter 9. Validation testing methods are applied in this stage for both systems to test system usability and query performance in term of accuracy. The informatics evaluation framework is used to evaluate the G3DMS design goal.

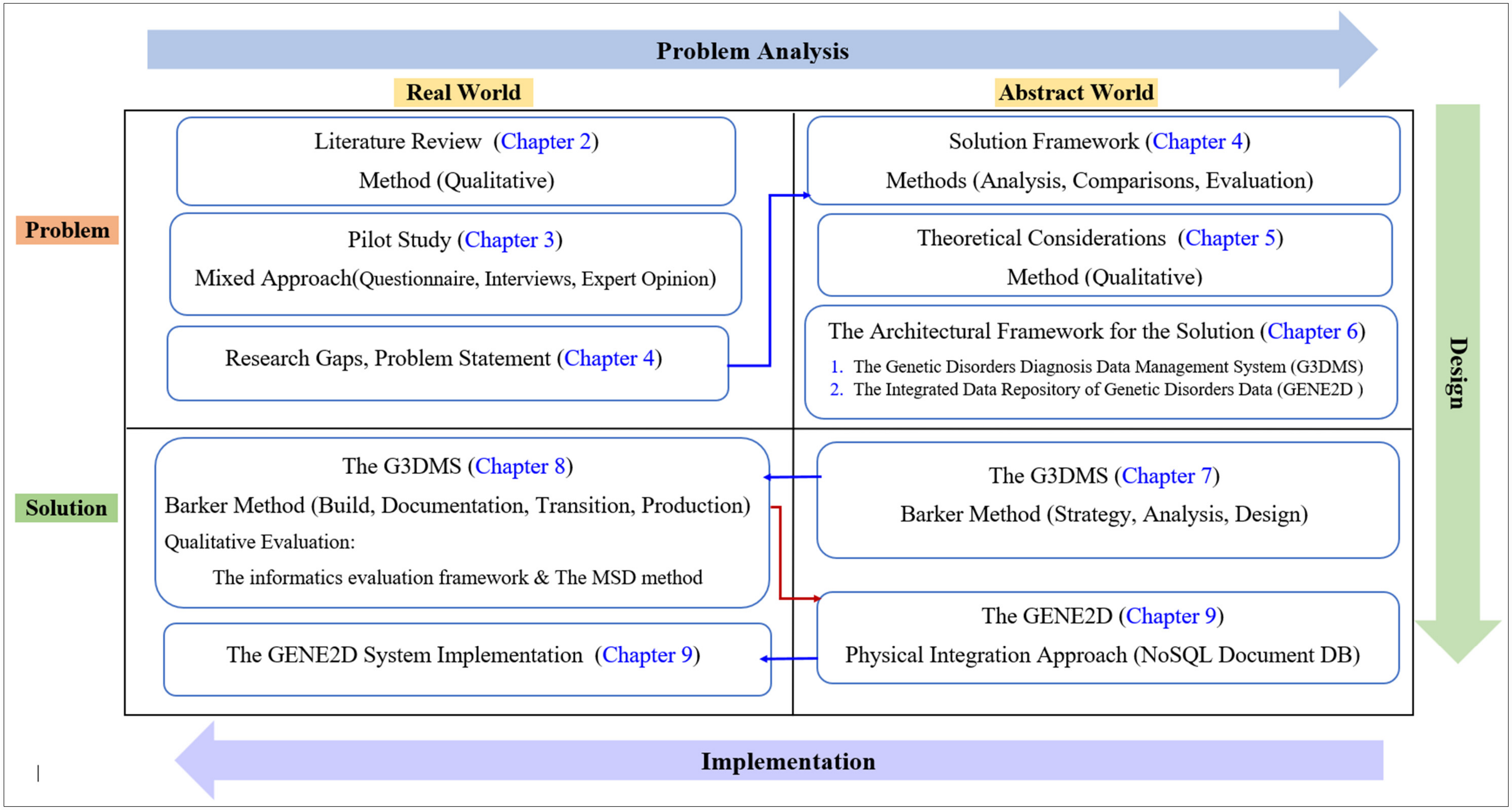

Problem Analysis
Real World
Abstract World
Problem
Solution
Design
Implementation
Literature Review (Chapter 2)
Method (Qualitative)
Pilot Study (Chapter 3)
Mixed Approach(Questionnaire, Interviews, Expert Opinion)
Research Gaps, Problem Statement (Chapter 4)
Solution Framework (Chapter 4)
Methods (Analysis, Comparisons, Evaluation)
Theoretical Considerations (Chapter 5)
Method (Qualitative)
The Architectural Framework for the Solution (Chapter 6)
1. The Genetic Disorders Diagnosis Data Management System (G3DMS)
2. The Integrated Data Repository of Genetic Disorders Data (GENE2D)
The G3DMS (Chapter 8)
Barker Method (Build, Documentation, Transition, Production)
Qualitative Evaluation:
The informatics evaluation framework & The MSD method
The GENE2D System Implementation (Chapter 9)
The G3DMS (Chapter 7)
Barker Method (Strategy, Analysis, Design)
The GENE2D (Chapter 9)
Physical Integration Approach (NoSQL Document DB)

Figure 1.2: Overall research methodology

1.6 Thesis contributions

This section provides an overview of the thesis results and the objective to contribute to knowledge in the area of health informatics (the use of data management systems and data integration repositories in healthcare and research), as well as to contribute to the R&D community in Saudi Arabia and the global community.

1.6.1 Current use of health information systems in research in Saudi Arabia

I. Participation in Saudi literature

The preliminary work of this thesis involved conducting a comprehensive examination of the health informatics state in Saudi Arabia based on the available literature. The goal was to identify the extent of application of health information systems within the framework of health informatics, in other words, the extent to which health information systems are used during routine care in addition to enabling secondary use of data in support of clinical decision making and medical research. This review identified the problems facing the application of health information systems and identified the reasons behind the delay of the full implementation of health information systems in all health organisations, especially small ones. The main contribution of this review was in its analysis and suggestions for successful implementation, citing successful experiences from developing and developed countries. The review also revealed some critical research gaps in the scope of interest that were not covered in the literature.

II. Pilot study

The study identified the major issues that hindered the use of healthcare data sources such as electronic health and medical records for medical research. A mixed-methods study of a questionnaire, interviews and expert opinion was used to examine and analyse the current state of clinical and genetic data available for medical research in Saudi hospitals, clinics and research centres. The study identified significant obstacles in term of data collection, storage and management for research. The main contribution of this study is introducing possible solutions to enhance the role of health information systems in medical research in Saudi Arabia.

III. Identify research gaps and propose a solution

The results of both the literature review and the preliminary study helped to set a solid foundation for the thesis, which is an effective contribution to solving the problem. Based on the gaps identified in research in the area of interest of genetic disorders diagnosis data management and data provision for research, this thesis bridges these gaps and proposes a solution. There are two research contributions: a novel data management system for the diagnosis of genetic disorders called G3DMS, which is suitable for use in any genetic clinic and research centre; and a NoSQL-based integrated data repository for genetic disorders data called GENE2D to aggregate genetic conditions data from multiple sources of genetic clinics or research centres.

1.6.2 Data management system for the diagnosis of genetic disorders

A case study was undertaken to analyse health information systems in Saudi Arabia to understand design problems via a brainstorming method with a focus group from the Princess Al-Jawhara Centre of Excellence in Research of Hereditary Disorders, King Abdulaziz University Hospital, Jeddah, Saudi Arabia. Barker's system design method and the multilevel service design approach were used together to guide the development lifecycle and evaluate each step in an iterative process. Also, a prototype was used to validate the proposed system via usability testing and a health informatics validation framework. The G3DMS comprises electronic data capture forms for data entry; a customised query builder to display and modify patient data as well as form research queries; a module that allows historical data to be uploaded in the form of bulk data using a template; export data options to Excel and JavaScript Object Notation format; and authorisation access for healthcare researchers and clinicians. The G3DMS is implemented in the Princess Al-Jawhara Centre of Excellence in Research of Hereditary Disorders (PACER-HD) at King Abdulaziz University Hospital. The major contribution of the G3DMS is its novelty, health informatics compliance, cost-effectiveness and ease of implementation in a low-resource setting, and customisable query interface for use in research.

1.6.3 Integration framework for genetic disorders data

The thesis incorporates an integration framework into the G3DMS using an integrated data repository based on the NoSQL database. The design of the NoSQL-based integrated data repository of genetic disorders data (GENE2D) includes three major components: the data sources, the integrated data repository as a central database, and the application interface. The novelty of GENE2D is in embracing a physical integration approach valid for a limited ICT infrastructure environment, using a NoSQL document store via MongoDB as a backend database for the integrated data repository, and in its application interface called Query Builder which provides customised query services to answer simple or complex research questions. The GENE2D demonstrates its potential contribution to benefit genetic studies and public health research and to help establish and grow a national genetic disorders database in Saudi Arabia.

1.7 Thesis structure

The thesis is organised into ten chapters. The first chapter provides the introduction and background to the problem, overview of the problem in the Saudi context, the research motivation, the research objectives and aims, the thesis approach and methodology, and thesis contributions. The remainder of the thesis is organised according to the objectives in Section 1.5.

Chapter 2 presents background information on the problem domain and examines relevant materials in the published literature in the Saudi context to identify the current situation and detect possible gaps in research.

Chapter 3 demonstrates a preliminary problem identification using a pilot study using mixed methods of a questionnaire, interviews and expert opinion. Results analysis is used to identify the study findings.

Chapter 4 presents the problem definition based on discussing research gaps identified from the literature and the findings from the pilot study. It formulates the problem statement and uses research questions to facilitate the steps toward the framework of the solution.

Chapter 5 provides the theoretical background information related to the design and implementation of a data management system and databases.

Chapter 6 demonstrates the proposed architectural framework of the solution: a standalone genetic disorders diagnosis data management system called G3DMS, and a NoSQL-based integrated data repository of genetic disorders data called GENE2D.

Chapter 7 reports the first three phases of the design lifecycle of the G3DMS: strategy, analysis and design.

Chapter 8 reports the next four phases of the design lifecycle of the G3DMS: build, documentation, transition and production.

Chapter 9 presents the G3DMS integration framework: the design and application of the NoSQL-based integrated data repository of genetic disorders data (GENE2D) from multiple sources (G3DMS).

Chapter 10 summarises the main thesis findings and identifies future research directions. It highlights significant design methods and presents thesis outcomes and contributions. Finally, it presents the thesis results and outlines the expected contribution to local and global efforts in future works.

1.8 Summary

This thesis examines the application of a health informatics framework in health information systems in Saudi Arabia to determine the problem domain and detect gaps in research from the literature and the pilot study. The thesis solution addresses the gaps and contributes to the existing body of knowledge and enhances the genetic conditions diagnosis process and research outcomes in the area of genetic disorders and public health. The thesis provides a solution with the following specifications: (i) cost-effective design based on new efficient technology and open source software; (ii) applicable for low-resource settings with limited ICT infrastructure; (iii) health informatics compliance framework; and (iv) customisable query interfaces to serve simple and advanced research studies at the individual clinic and centre and national levels.

The thesis proposes a new data management system, G3DMS, which meets the health informatics framework by supporting diagnosis workflows, diagnosis decisions and the reuse of data for research. In addition, this research is considered one of the pioneers in using a NoSQL document database as a storage mechanism for an integrated data repository. The flexible schema design of GENE2D can accommodate heterogeneous clinical and genetic data with various structures and formats. Therefore, the contribution of the thesis outcomes is not only relevant to current challenges, but it engages in the emerging field of big data integration because it adopts NoSQL technology in the physical integration approach.

The following chapter, Chapter 2, presents the first step in investigating the problem in the real world, in a literature review of the area of e-health application and health informatics status in Saudi Arabia.

Chapter 2: Literature Review and Background

2.0 Chapter overview

This chapter reviews the field of health information systems and e-health application in Saudi Arabia and identifies research gaps. This chapter has the following structure. Section 2.1 introduces the field of health information systems. Section 2.2 presents an overview of health information systems regarding clinical data management, data integration, and secondary use of clinical data in research. Section 2.3 covers the significance of health informatics in healthcare management. Section 2.4 provides background information on healthcare delivery in Saudi Arabia, the application of health information systems in Saudi hospitals along with implementation barriers, e-health and health informatics developments, and the status of medical research performance in Saudi hospitals. Section 2.5 discusses research work in the area of health information systems, e-health and health informatics. Section 2.6 presents a summary and the implications of the identified gaps in the literature review.

2.1 Introduction

This thesis focuses on the use of healthcare data to improve the care process, assist clinical decisions and enhance research studies. The research approach and methodology were presented in Chapter 1, which includes research questions structured to guide the thesis methodology through the three development stages of analysis, design and implementation (see Figure 1.1 and Figure 1.2). This chapter is the first step in the problem analysis stage in the real world to answer the research question: *What is the current state of e-health and health informatics in Saudi Arabia?* The main objective is to cover all aspects of the current status of health information technology implementation in Saudi Arabia with a focus on the use of health information systems for research. The results of this chapter are used to identify the problem domain and detect possible research gaps.

Preliminary

Healthcare delivery generates a large amount of data during the routine care process. Paper-based health records are the widespread, conventional form of keeping patient data which can be hard to manage and protect from being misused. However, patient data in these records exist in such a scattered and inaccessible state that it prevents their contribution to growing knowledge and fails to support clinical processes at the point of care and population health [26]. Digitising healthcare services can efficiently store, process and transmit enormous amounts of data which will improve healthcare quality and safety at a lower cost [27]. The use of IT-enabled healthcare applications such as electronic medical records and computer-based patient records have facilitated healthcare processes. However, the increasing complexity of healthcare procedures and the involvement of multiple healthcare providers in inpatient treatments necessitated more IT involvement for integrated care [28]. Information and communication technology (ICT) played a significant role in health information systems reform which has resulted in the acceptance of health information technology in healthcare systems [29]. Electronic health records provide an integrative view of a complete record of clinical patient visits, as well as supporting other care-related activities [28]. The application of ICT for health, known as e-health, is one of the primary tools for healthcare delivery, health literature, and health education, knowledge and research [30]. E-health and health informatics are related terms for all forms of electronic processes in healthcare which are provided through vast channels of ICT [31], [32]. Poor ICT mostly results in undesired implementation outcomes for e-health [33]. Well-designed and implemented e-health may have a major impact, especially in under-resourced settings that lack adequate infrastructure and backup systems, as is the case in developing countries [34]. However, variations in the outcomes of a particular implementation arise directly from the inconsistency between system design assumptions and the working environment [26]. Saudi Arabia has set, through its Ministry of Health, a national e-health strategy aiming at leveraging healthcare delivery by improving standards, availability, quality and equality throughout the country [35]. In 2011, a five-year plan was initiated to enable the vision and serve goals to care for patients, connect health providers at all levels, measure healthcare delivery performance, and transform healthcare delivery to a world-class standard [36]. However, the health sector's dependability on the new technology varies in terms of the extent and the level of embracing e-health and the adoption of health information systems such as electronic health record systems. Barriers and challenges in embracing such a technology for healthcare services in Saudi hospitals arise during the implementation, maintenance and improvement phases [37]. The lack of a health informatics framework in Saudi Arabia due to the deficiency in health information management, communication and sharing, and lack of a standard format for data definition, storage, structure and organisation, prevented the establishment of a national shared electronic health record for a national health information network [38].

2.2 Overview of health information systems

A health information system or clinical information system is a long-term automated database system that contains clinical information used for patient care [39]. Other terms are used to refer to a health information system such as health information technology, clinical information system, health information management, hospital information systems, healthcare information systems, and health application systems [40]. In this thesis, the term health information system (HIS) is used interchangeably with hospital information systems where appropriate to the context.

The health information system function is focused on capturing, storing, processing and delivering information to clinical decision-makers. The health information system is considered an important component of hospital information system solutions, such as the electronic medical records system, computerised physician order entry system, laboratory information system, pharmacy information system and radiology information system [8]. The use of health information systems, such as electronic medical records in healthcare, can improve healthcare quality, prevent medical errors, increase administrative efficiencies, reduce paperwork, and expand access to affordable care [41]. Clinical data often resides in inaccessible repositories making linkage and data sharing as well as data analysis and interpretation challenging for a patient or population-level research studies. In addition, the clinical data in a traditional electronic medical record is generally not available to a researcher from outside the organisation. On the other hand, the emerging electronic health record in the care setting comprises a digitised version of all clinical data relevant to patients' care such as demographics, medical history, medication, care plan, laboratory data, radiology reports, physicians' feedback and billing information [3]. It stores patient information with full interoperability between systems in a health organisation, so all services provided to the patient are stored in the patient record [42]. Although an electronic medical record is a patient-oriented system and provides a full view of a single individual case, it lacks the ability to integrate a group of patients with a common condition to support population-level studies. Also, the patient record may contain scanned documents and reports that are difficult to search for content [43]. It is also difficult to use this valuable tool for research and analytics unless it is joined in a single repository to allow more insight and use historical data for identification of more risk factors and patients at risk of some chronic diseases [44]. Furthermore, the inconsistency of the clinical data platform leads to a significant reduction in the prevalence of electronic health records and its use to develop knowledge [2]. Hence, for electronic health records to be used as a source for medical research such as observational studies in registries, standardised terminologies will need to be developed to be used across the platform for better analysis [45]. Clinical data are heterogeneous in nature and usually stored in multiple clinical and operational systems, as the same clinical information may be stored in multiple forms and formats across health information systems making the data management process and integration difficult [46].

2.2.1 Clinical data management challenges

Clinical data is composed of information that includes determinants of health such as biomedical, demographic and genetic factors; health behaviours; socioeconomic and environmental factors; measures of health and health status, such as laboratory data, physical examination results, imaging studies, diagnoses, established treatments and responses to applied interventions, and documentation of care delivery. These data captured during the clinical care delivery, administration and claims, or research processes are mainly stored in several databases distributed across the healthcare system. In general, the data collected through clinical practice software tools include administrative and demographic information, diagnosis, treatment, drugs, laboratory tests, physiological surveillance data, and health insurance. Operational systems with relational databases are used to manage patient-level data, and healthcare businesses, whereas data warehouses and data marts are used within a single organisation for reporting, use online analytical processing systems. Over the past decades, healthcare organisations gathered patient data into their databases to make more informed clinical decisions. Often data are stored in a relational database, such as Oracle, SQL Server and DB2 [47]. Many healthcare organisations have problems offering a comprehensive view to their researchers as data are stored separately in different locations, often under different IDs [48]. The current method used to collect and store clinical data makes data management for researchers very daunting and makes institutional data exchange and data sharing among healthcare organisations very difficult. On the other hand, having data in a central location allows comprehensive data analytics that can effectively contribute to different areas such as clinical operations, research and development, and public health to provide a better outcome and reduce waste and inefficiency [49]. Using various data analysis techniques and mechanisms to harness the best value and better healthcare outcomes from medical data, repositories have the potential to advance the healthcare industry [50].

2.2.2 Clinical data integration and sharing challenges

Health information systems for any health organisation comprise different software solutions for data collection, storage and management. Data reside in different locations within the same organisation in various formats and structures, making data manipulation a complicated task. Lack of medical records integration among an organisation's physicians or across institutions creates great difficulties in data analysis and medical research studies [51]. In addition, current patient clinical care databases are often inadequate to assess health interventions because data are often missing or incorrect, and it is difficult to link patients' demographics and clinical care procedures [17]. Therefore, healthcare institutions need fundamental changes in the infrastructure and mechanisms for data collection, storage and exchange. For this data to be suitable for use in research, systems must have the ability to be integrated to create a comprehensive view and provide meaningful insights for patients and researchers. But it is very challenging to integrate databases with different specifications in data models, database schemas, the queries they support and the terminologies they use. Data sharing through databases is more common practice for clinical research as data collected at multiple sites are integrated with some disease-oriented database systems, since one location may not be able to provide sufficient data for analysis. Clinical institutions may also be limited in terms of research interests, so a common database can make the collected data available to researchers in a variety of locations [14]. The process of data integration requires bringing together scientific methods and specifications needed to be stored in a database. The overall goal of data integration for the clinical research community is to be able to answer questions about aggregated data which can be very difficult if each individual data source must be accessed separately or sequentially.

Data integration options

Data integration is a prerequisite for obtaining a unified view of clinical and genetic data from multiple operational data stores for different healthcare organisations. Integration of data stored in heterogeneous database management systems can be achieved by aggregating across data sources using an intermediate solution to retrieve data or combine copies of data in a centralised system from multiple sources to provide a unified view using various techniques for data extraction. The process of data integration involves extracting, transforming and cleansing the data before aggregation. Several approaches have been used to integrate data from different sources using warehousing approaches (combining a copy of the data in a new repository for further analysis) or using mediation-based approaches (applying conceptual schemas to bridge representational heterogeneity of the databases and providing queries with the ability to collect and integrate data from distributed sources). Other techniques also applied to help integration include semantic data modelling, ontology definition, query translation, query optimisation and terminology mapping [18].

There are numerous integration efforts such as using approaches to integrate terminologies, ontologies and schema matching [52]. Integration solutions such as data warehouses that require a global schema find it difficult to handle diverse sources with independent requirements or extremely complex schema to encompass all data requirements [53]. Therefore, researchers face a significant barrier of insufficient technical infrastructure to support data collection manipulation and proper analytical tools. An integrated data repository with the traditional data warehouse structure and a predefined data model incorporated into the database schema constitutes the major source for the data required for clinical and translational research, designed to integrate various aspects of patient care data [54]. The design of integrated data repositories for medical data must consider critical issues such as identity management, protection of confidentiality, convenient and flexible queries, and semantic and ontology definitions at the schema level for data from different data sources with different naming and structural conflicts [55]. Data organisation and representation in integrated data repositories is very important for analysis, as they accommodate data from a range of sources which may require a continuing extension. Therefore, the extension should be allowed without any major changes in the original schema to ensure easy maintenance and data analysis [56].

2.2.3 Secondary use of clinical data for research challenges

The delivery of healthcare in the health sector results in a large amount of patient information that is usually spread throughout an organisation’s health information systems. Some data are administrative and claim data, and others are clinical data. Reuse of clinical data is essential to deliver quality healthcare, improve healthcare management, reduce healthcare costs, manage population health, and maintain efficient clinical research [57]. Despite the many potential benefits of using health information systems to improve the quality, safety and efficiency of healthcare, reuse of clinical data faces multiple challenges in terms of the purpose of collection, data quality issues, data integration and interoperability limitations, and organisations’ culture [58]. For example, the secondary use of electronic medical record data has intolerable limitations in the system structure that only provides a single level view of an individual patient and insufficient aggregation to present a group level that limits a population’s health research requirements [59]. The quality of data collected through health information systems such as electronic health records is often questioned in relation to data reliability and usefulness for research applications [12]. There are many reasons for poor data quality in these systems, such as incomplete, inconsistent and noisy data. Sometimes, physicians collect their findings using free-text notes, or by dictation, and these reports need to be transcribed into the computer [13]. In addition, there are legal and ethical considerations in secondary use of healthcare data especially in the presence of patient identification data which mostly depends on the organisation culture for data governance and whether patient consent is obtained for data information reuse [57]. Clinical data reuse typically involves combining heterogeneous and multidimensional datasets into shared data repositories, data warehouses or networks, with challenges of integration, interoperability and shared meaning [58].

2.3 Health informatics framework: an overview

This thesis focuses on health informatics from the perspective of technology support to enable information to be collected, processed, transformed and shared to support care delivery, decision making, and reuse in medical and public health research. Therefore, this thesis adopts the definition presented by Coiera et al.: "health informatics is the study of information and communication processes and systems in healthcare" [26]. The terms e-health and health informatics are both used to refer to all forms of electronic health services and information delivered or enhanced through ICT to improve health and the healthcare system [60]. The role of health informatics is to develop healthcare systems that develop the organisational processes and structures to facilitate the method of collecting, communicating and applying clinical evidence to routine care [26]. Health information technology has contributed to improving healthcare management and delivery by using powerful and cost-effective technological methods for collecting, storing and retrieving information [61]. In general, health informatics is an interdisciplinary field that aims to optimise the use of information to influence healthcare services and outcomes by using health information technology. Health information systems should be designed with health informatics principles to be effective and serve the role of using the information in the delivery of care and for research [62].

2.3.1 Patient care delivery support

Health information systems help to digitally document each event of a patient visit or seek treatment and receive care or referral for care from a healthcare institution [63]. The integration of the information flow between electronic health records and other health information systems such as laboratory and pharmacy information systems enhances the speed of the data processing, which optimises the diagnosis workflow and improves the healthcare process [63].

2.3.2 Decision support

The use of a health information system such as electronic health records prevents errors which may result from using handwriting to capture data during the care process [64]. In addition, electronic health records provide healthcare information in an organised, standardised and encrypted format so that this information can be managed to assist healthcare professionals and patients through decision support and information linkage and coordination. Data captured at one site from one authorised source can be shared in other sites allowing reuse of data in the process of care. For example, data entered by a physician about drug prescription can be used for a care plan, billing system and so on [63]. Physicians gaining easier access to data supports them in adopting the best diagnostic and treatment strategy [63].

2.3.3 Secondary reuse of patient data

Digitisation of patient care data facilitates the secondary reuse of such data in clinical research or public health. However, a strict policy is required to protect health data security and patient confidentiality providing authorised levels of access to patients and health professionals. Security and confidentiality rules can be enforced for the reuse of patient data by implementing new infrastructures such as clinical data warehouse and IT platforms dedicated to translational research [63].

2.4 Background on the healthcare system in Saudi Arabia

The establishment of the healthcare system in Saudi Arabia coincided with the expansion of the Kingdom of Saudi Arabia by the founder, King Abdul Aziz Al Saud. Public health and disease control were among the top priorities for the government. In line with the founder's vision for modern national healthcare services, the government's focus has been on building a strong infrastructure for a comprehensive health sector to serve the entire country. In 1951, the Ministry of Health was established with a clear mission to deliver healthcare at all levels, promote public health and disease prevention, and develop laws and regulations governing the public and private health sectors [65].

Over the past decades, the government has established essential infrastructure for primary healthcare, hospitals and research facilities with the primary objective of providing a free healthcare service for all citizens, which represented a "Health for All" goal [66]. Health services in Saudi Arabia are delivered through multiple channels, with the Ministry of Health being the main provider managing almost 60% of services [67]. However, approximately 20% of services are provided by government agencies to serve certain divisions of the Saudi population such as the Medical Services of the Armed Forces, the King Faisal Specialist Hospital and Research Centre, the Ministry of National Guard Health Affairs, security forces hospitals, and university hospitals; in addition to the private sector that provides the remaining 20% of healthcare services [42], [32]. The Ministry of Health is authorised by the government for administration, planning, funding and control, and is responsible for the overall supervision of healthcare facilities in both the public and private sectors [32]. The Ministry of Health has followed a five-year development plan since 1970, which is part of the overall development of the country [68]. According to the Millennium Development Goals report in 2013, Saudi Arabia made extraordinary improvements to health services, especially in the field of primary healthcare in terms of efficiency and level of integration [69].

2.4.1 Health information system application in Saudi Arabia

The demand for healthcare services is increasing as a result of the high rate of population growth, the current lifestyle that increases the likelihood of non-communicable disease risk factors, and the vast area of the country and the distribution of people in a remote area, which are all challenges for healthcare delivery [68]. As a result, in 2008 the Ministry of Health promoted a national healthcare strategy to support a nation-wide transition to the use of electronic medical records systems in the health sector to improve data quality and progress towards paperless management [67]. In addition, health providers in Saudi Arabia must also try to deal with the increasing need for information according to their resources and strategic plans. Therefore, the Ministry of Health adopted a health information system model (Health Information Centre) comprising the Department of Computer Science and the Department of Statistics to conduct data collection, processing and analysis, and provide all the information required to help with decision making. The strategy for the National Health Information Network was proposed in 1996 with the aim of establishing a standardised network infrastructure to enable healthcare organisations to communicate nationally and share information for improved quality care, reduced the operation cost and enhanced access to patients' databases [70]. Over the past three decades, there have been initiatives from the Ministry of Health to implement health information systems such as electronic medical and health records systems nationwide [71]. The application of health information systems in hospitals throughout the country varies, as some hospitals lack computerised systems [72], and other hospitals use systems from different vendors [32]. For example, the adoption rate of electronic health and medical records in Saudi Arabia has grown at different levels according to the hospital size, type and ownership across all regions with multiple challenges that delayed full implementation [37]. Successful implementation cases in e-health, particularly electronic health and medical records, included King Faisal Specialist Hospital and Research Centre; the Ministry of National Guard Health Affairs; and the security forces hospitals [42], [32]. The Saudi Arabia National Guard Health Affairs received the distinguished excellence in Electronic Health Records Award at the 2010 Arab Health Awards [73].

2.4.1.1 Health information system adoption challenges in Saudi Arabia

Despite the positive outcomes and the number of objectives and initiatives that can be achieved through the implementation of health information systems in Saudi Arabia, the adoption pace of systems such as electronic health and medical records remains slow and faces numerous challenges and barriers that significantly influence the contribution to healthcare delivery [4]. Deficiencies in the degree of completion of implementation due to the legacy of paper systems conversion as well as poor maintenance of the system software updates and lack of interoperability and information exchange all led to variation in the adoption level among hospitals [37]. Underutilisation of electronic health records functionalities among hospitals is also considered one of the barriers to adoption and to the successful implementation of these systems [74]. In addition, issues related to data capturing, data sharing and lack of health informatics professionals are reasons for unsuccessful implementation and utilisation of health information systems and electronic patient records [75]. Previous research into the application of health information systems in Saudi hospitals identified several barriers to their successful implementation which can be summarised into the most cited categories: technical barriers, human/behaviour barriers, organisation barriers, and financial barriers.

Technical barriers

Technical challenges related to poor infrastructure and system efficiency (hardware and software) might complicate the implementation of systems [76], [77]. Vendors’ instability is one of the most recognised barriers to health information systems; it is difficult for system users to adjust to the frequently changing software packages [78], [79]. Complexity in system software design, the health data standards adopted in these systems, and lack of system customizability are considered great barriers to mastering these sophisticated systems [5], [80]. Lack of systems interoperability which prevents data integration is one of the barriers to successful adoption [79]. Security concerns regarding the use and access to electronic health record systems are considered a threatening barrier to the implementation [79], [81], [76]. Lack of backup plans for system maintenance and downtime is one of the strongest barriers to the adoption of the electronic health records in Saudi Arabia [82], [79], [81].

Human barriers

User awareness and satisfaction with the use of health information systems is a significant factor in the successful implementation and adoption of the systems [82]. Lack of computer literacy skills and English language mastery that influence knowledge about systems constitutes a major barrier to system adoption [81], [78]. Lack of continuous training and support by IT staff leading to an increased workload during data entry is also a critical barrier to implementing systems in Saudi hospitals [77], [80]. User resistance to new systems due to technical issues, and the system not fulfilling its purpose or not meeting users' expectations are social barriers influencing the adoption of systems in Saudi Arabia [76].

Organisational barriers

User training and technical support are among the organizational factors affecting the adoption of health systems in Saudi Arabia, in addition to other barriers such as:

- Lack of end-user training and technical support [76].
- Organisational cultural issues, bureaucracy and human resource issues [32].
- Not meeting privacy and security standards [81].
- Lack of management experience influenced the selection of the software of electronic medical records [76].
- Unpreparedness to change and the lack of redesigning the workflow to match the electronic health records [76].
- Lack of strategic planning for the adoption and implementation of electronic health records [4].

Financial barriers

Financial resources are determined to be a significant factor in the transition from traditional paper medical records to electronic health records [83]. Financial cost related to money and funding is one of the most commonly cited barriers to the adoption of electronic health records systems [81]. For example, lack of capital resources to invest in electronic health records, the high initial cost of implementation, and high operation and maintenance costs are among the most perceived financial barriers in Saudi Arabia [4].

2.4.2 E-health and health informatics development in Saudi Arabia

Health services are provided in Saudi Arabia through multiple providers with proprietary systems from different vendors [37]. Therefore, e-health services in the country differ from one organisation to another with no standardisation of health information systems, as there are no specific methods for how data can be predefined, characterised, organised, stored, exchanged, integrated, accessed and controlled [38]. These variations result in a lack of connectivity and interoperability between systems, which is the main reason for e-health application failure and delay in establishing a national electronic health record [84]. As well as other barriers hindering e-health use in the country such as lack of medication safety, privacy and confidentiality concerns, culture and human barriers [83]. In the year 2000, the health reform committee mandated by the government to review healthcare services highlighted the lack of appropriate health informatics applications as one of the main challenges in the healthcare sector. The Saudi Association for Health Informatics (SAHI) was established in 2005 as a result of a need for proper health informatics applications in the health sector [84]. The role of SAHI is to promote health informatics knowledge by organising scientific events such as the e-health conference which is held roughly every two years in Riyadh, Saudi Arabia [38]. In support of this trend, King Saud University for Health Science took the initiative to design a health informatics Masters program to prepare health informatics specialists who would be able to participate in research in the health informatics field [84]. The Saudi Health Informatics Competency Framework was developed to define the field in Saudi Arabia and set the boundaries to distinguish it from areas such as bioinformatics based on the competencies outlined in the framework [85]. The development of health informatics in Saudi Arabia has the potential to achieve sustainable healthcare systems using advanced health information technologies to deliver quality service and enhance medical and public health research [38].

2.4.3 Medical research in Saudi Arabia

Medical research in Saudi Arabia is almost entirely confined to medical schools and their affiliated teaching hospitals that provide medical education, training and research, in addition to delivering primary and specialised level care [70]. In Saudi Arabia, hospital libraries traditionally served as a primary source for medical research and information required to improve the healthcare delivery system, especially in teaching hospitals [86]. A number of studies have addressed issues and limitations of healthcare libraries in the provision of information services and the use of information and communication technologies. A design of a Saudi Health Information Network was proposed by Khudair in 2005 with the aim to link all appropriate health sites and provide various collaboration and communication channels for health professionals and health information professionals [87].

Hospital databases are a key component of healthcare management and productive medical research, especially in university hospitals in Saudi Arabia as these databases are deemed to be a major source for researchers to conduct observational or clinical studies. The source of hospital databases is patients' data collected from hospital information systems, electronic medical records or computerised physician order entry systems. Despite being a rich source and providing a promising future for medical research, adaptation rates for health information systems and electronic medical records in Saudi Arabia are still low due to human, technical and financial barriers [88], [4].

Disease registries in Saudi Arabia provide a comprehensive data source which demonstrates the occurrence of health problems and determine the risk factors for planning preventive strategies. The Department of Saudi Diseases Registries aims to establish a national database for health-related events and diseases, produce national health statistics, and provide support for clinical and epidemiological research [89]. Several disease registries are available in Saudi Arabia such as the National Cancer Registry enabling researchers to analyse the prevalence and other determinants of various diseases according to the registry scope and geographic coverage. However, data in registries are mainly affected by case definition, and researchers need to be aware of changes in case definition overtime before including or interpreting the information. Data accessibility in most registries is limited to the annual reports [90]. A few studies presented web-based designs for disease registries such as the Saudi National Diabetes Registry and Congenital Glaucoma Registry which can only be accessed with unique login identification and password, with an electronic case report form used to gather sufficient information from patients with extra validation checks to ensure data quality [91], [92].

Other potential sources for producing specific healthcare data in their topic of interest are scientific societies, associations and centres which can contribute to health research by the actual data or by providing a reference to related studies in the area [90]. Most medical research centres are affiliated with universities and their teaching hospitals, for example, King Fahd Medical Research Centre, King Faisal Specialist Hospital and Research Centre and King Abdullah International Medical Research Centre which depend on their own research databases and disease registries [93]. But studies have found that hospitals and research centres in Saudi Arabia had made a very limited contribution between 2008-2012 to international medical publications [94].

2.5 Discussion

The expected role of health information systems is to increase the efficiency of healthcare delivery within health informatics principles by assisting in the clinical workflow, providing support for clinical decisions, and allowing data exchange and reuse for research [63]. However, the experience in Saudi healthcare systems varies between successful implementation in some hospitals and failure to adopt such a system in others [37]. It is crucial to identify the areas of weaknesses and barriers in order to achieve efficient implementation and full use of the health information system functionalities [74].

Most of the studies in the Saudi context present the current state of health information system challenges and opportunities and make recommendations from lessons learned and other countries' experiences [95]. The majority of the studies report various implementation experience in electronic health and medical records systems adoption in different large-sized Ministry of Health public hospitals located in major cities or a group of regional hospitals in a certain division in Saudi Arabia. These studies were limited to presenting the factors that influence system implementation based on field research or literature review analysis of quantitative, qualitative or mixed-method approaches, then providing some recommendations to overcome barriers to implementation.

A few studies presented a theoretical framework using modern technologies to model a solution for some specific problems. Almuayqil et al. proposed an integrated framework of knowledge management and knowledge discovery to help overcome e-health barriers [96]. Al-Shehri investigated the role of health informatics in transforming public health services, research and education in Saudi Arabia, then proposed a public health informatics framework to be hosted by the Council of Health Services to facilitate the linkage and integration of current public health data and information available at different health providers to prepare for a national public health informatics system over time [97]. In the area of cloud computing decision making in the healthcare sector, Alharbi et al. proposed a strategic framework focused on five dimensions: organisation, technology, environment, human and business [98]. A cloud-based conceptual framework was suggested by Kurdi et al. for e-health systems with the aim to create centralised hospital databases in Saudi Arabia, and the study presented the system architecture and described the functionality of different modules [99].

In addition, several PhD studies in this area have investigated the implementation of health information systems, on a larger scale and provide solutions. Alanazy conducted a study to explore the implementation of electronic health record systems, together with the difficulties associated with their application in Saudi hospitals, and identified several factors to be significant barriers such as cost, privacy and security concerns, software complexity, vendor instability, ongoing maintenance, and lack of support for uniform standards in the various software packages [79]. A similar study by Alghamdi investigated factors associated with the implementation and adoption of electronic health records and pinpointed the actual barriers to their adoption in Saudi hospitals, such as lack of user experience, security concerns of using systems, resistance to technologies, and high cost of adoption [81]. Hasanain also attempted to provide a solution to the variation of electronic medical record implementation by developing an implementation framework to guide implementation in Saudi public hospitals [82]. In addition, Sabbagh presented a novel model for managing health informatics to address the issues identified in the literature and the field study; the model was evaluated in Saudi private health organisations, and a health informatics management model was developed in three iterative development and evolution stages to reach the final model [100]. Moreover, a study by Alzghaibi to explore the large-scale implementation of electronic health record systems in primary healthcare centres in Saudi Arabia found the main barriers behind the implementation failure were the scale of the project, shortage in health informatics expertise, lack of training and support, end-user involvement, software selection, and geographic challenges [76].

The literature on the Saudi context shows that there are few sources to obtain clinical and genetic data for use in research. Yet, these sources are inefficient either because of the difficulty in accessing them or for using traditional methods of data collection, which makes them invalid for use in research. Limited research studies on the challenges facing researchers focused on factors affecting researchers in conducting clinical research in particular clinical trial studies. Researchers' interest in conducting research, the long process of research approval, inadequate time, and financial funding are among the most frequent challenges reported by most authors [101], [102], [103]. Another study by Sheblaq and Al Najjar presented other barriers such as insufficient training in carrying out clinical research and lack of designed system operating procedures for the research process in the Arab region; they also emphasised that there is a great need to provide a database or website for the research project available for interested researchers [104]. Al Dalbhi et al. conducted a study to assess the methodological difficulties encountered by healthcare practitioners who have conducted or have been involved in any kind of clinical research in Saudi Arabia and showed that receiving funds and financial resources are the greatest challenges besides biostatistician availability [105]. In addition, Ali et al. reviewed clinical trial activities in Saudi Arabia over 15 years and reported that the lengthy ethical review process, difficulty in recruiting study subjects, and insufficient financial compensation were the reasons for the low level of activities [106].

2.6 Summary

Most research studies showed the barriers behind the variation of the implementation of health information systems among Saudi hospitals. Some efforts have been made to identify the cause of the barriers and the reasons behind the failure of implemented systems. However, few studies attempted to provide some solution frameworks for successful implementation of such systems in Saudi environments. To our knowledge, none of the studies showed a practical and applicable solution to leverage the healthcare processes and promote health informatics principles. In summary, this review has several implications. There is a need to conduct more research that focuses on IT solutions customised to fit the local health information technology infrastructure, which can be evaluated in the real, local healthcare environment. It is also essential to conduct more studies based on theoretical frameworks (models) to support the study hypotheses and provide the bases for evaluation. In addition, research studies should help present the role of health informatics application in Saudi healthcare systems and adopt health informatics frameworks for evaluating e-health services, such as the implementation of electronic health records in Saudi hospitals. Moreover, there is a need to undertake research to investigate the role of health information systems in medical research in Saudi Arabia from the perspective of healthcare professionals and IT specialists. Further investigation is essential in this area to understand the existence of the problem and its current state in Saudi Arabia.

In the following chapters, this thesis provides solutions to these gaps including a pilot study to research this area for further evidence to define the problem in Chapter 3, demonstration of the research gaps and the solution overview in Chapter 4, a theoretical consideration of the proposed solution in Chapter 5, the architecture framework of the solution in Chapter 6, the first three system design lifecycle phases of the genetic disorders diagnosis data management system called G3DMS in Chapter 7, the next four system design lifecycle phases of G3DMS in Chapter 8, and the development of GENE2D as the integrated data repository of genetic disorders data in Chapter 9.

Chapter 3: Preliminary Problem Identification Using a Pilot Study

3.0 Chapter overview

This chapter investigates the problem using a pilot study with a mixed-methods approach to examine and analyse the current state of clinical and genetic data available for medical research in Saudi hospitals, clinics and research centres. Section 3.1 highlights the essential objectives of the study, Section 3.2 illustrates the methods implemented in this study, including the survey questions, interviews and expert opinion, Section 3.3 explains the data analysis method applied, Section 3.4 presents the analysis of the results, and Section 3.5 presents the discussion of the results outcomes in three themes: data collection difficulties, data storage difficulties, and lack of system interoperability. Section 3.6 provides some recommendations. Section 3.7 concludes the findings of the study and narrows the scope of the thesis. Finally, Section 3.8 summarises the chapter key points and introduces the next chapter.

3.1 Introduction

This chapter forms the second step in the problem analysis stage in the real world (Figure 1.1, Figure 1.2) and provides the answer to the research question in the thesis approach, see Figure 1.1: *What are the challenges that a Saudi physician faces regarding the use of information from health information systems in medical research?*. The process of problem identification and definition began with the literature review in Chapter 2, which resulted in the identification of research gaps and recommended further investigation of the use of data from health information systems in medical research. In this chapter, the thesis bridges the gap and performs a preliminary pilot study to research this area for more knowledge and to be more informed. The main objective is to conduct a pilot study to investigate the current state of medical research in Saudi Arabia; identify possible issues that hinder the improvement of medical research and the role of hospital databases in supporting the research, and identify possible solutions to enhance the role of health information systems in medical research in Saudi Arabia. The results of this chapter are combined with Chapter 2 findings to formulate the problem statement and outline the overview of the solution.

Preliminary

Recent studies have shown hospitals and research centres in Saudi Arabia have made a very limited contribution to international medical research performance and publications. Although in recent times, the Saudi government has paid much attention to the adoption of hospital information systems and electronic medical records, the importance of using hospital information systems to enhance medical research has been neglected. Although medical research is tied to the provision of patients' data collected regularly from daily clinical observations and by sharing resources among hospitals, health organisations raise concerns about standardisation, data quality and security issues. Hence, inefficient databases that lack well-adopted standards impact research accuracy and quality outcome. The aim of this study is to perform field research in the use of hospital information systems in medical research and their role in forming efficient research databases, and use the findings to (i) support the research outcomes of this thesis in terms of its validity and enriching the knowledge in this area effectively; and (ii) narrow the scope of the thesis to a specific area of medical research databases to ensure an accurate source for the problem definition and propose a reliable solution.

3.2 Methods

The main objective of this study is to investigate the current state of medical research in Saudi Arabia and the role of hospital databases in enriching the research, analyse the ability of databases to be integrated using a standardised method for data collection, storage and retrieval, and identify challenges in achieving intended goals.

Based on the discussion and results of the preliminary review of literature in the field of medical research in Saudi Arabia, the decision was to perform more investigation and study the problem in the real world to gain a profound understanding of the capacity and efficiency of the current systems available for medical research and identify potential gaps in the literature. Therefore, the most appropriate strategy was considered to be a mixed-method approach to enhance the reliability and validity of the study and add credibility to the results. The field research study consisted of three components: (i) a survey that targeted all healthcare professionals who work in hospitals in Jeddah; (ii) structured interviews with IT professionals who were responsible for database management in these hospitals; and (iii) a formal discussion with a leading researcher in the field of genetic diseases and the founder of the Medical Genetic Unit at King Abdulaziz University in Jeddah, Dr Jumana Al-Aama, to acquire expert opinion on the use of clinical and genetic data for research related to genetic disorders.

The interview and survey questions were designed and developed at the same time. The medical research areas of hospitals covered in this study were biomedical, clinical/epidemiological and genomics. The research focus was to identify the key issues that obstructed the exploitation of healthcare data sources such as electronic medical and health records for medical research and highlight possible solutions to enhance the role of health information systems in medical research in Saudi Arabia and form the basis for the thesis proposed solution.

3.2.1 Survey

Participants

Survey participants (n = 503) were invited through heads of department in the following eight hospitals in Saudi Arabia: King Fahad General Hospital, Al Aziziyah Maternity and Children Hospital, Maternity and Children's Hospital, King Abdulaziz University Hospital, King Faisal Specialist Hospital & Research Centre (KFSHRC) and three private hospitals. These eight hospitals were selected on the basis of their provision of healthcare systems to support patient data management, and the likelihood academic physicians and hospital staff are required to perform medical research in the areas of biomedical, clinical/epidemiological and/or genomics. The first three hospitals listed are government hospitals (Ministry of Health non-teaching hospitals), while the following two are teaching hospitals involved in medical research. All eight hospitals agreed to participate in the survey. Department heads, who had ethical oversight of the survey in accordance with respective departmental codes of ethics, distributed the survey via a link in an email sent to all of their healthcare professionals.

Materials

An anonymous online survey with 32 close-ended questions was designed using Google Forms for this research to assess the current research challenges facing healthcare professionals at Saudi hospitals. It was designed to gather data from all healthcare professionals working in the Saudi hospital environment (see Q1 of Section 1 in Appendix 3.1). The survey was piloted through peer review and underwent ethical approval by the Ministry of Health. The questionnaire had 12 sections (see Figure 3.1) based on categories of items (e.g. Section 2: Research Contribution Part I and Part II in Appendix 3.1), where Part I determined whether subsequent items in Part II were relevant to participants, who could skip irrelevant questions. This can lead to more reliable results [107].

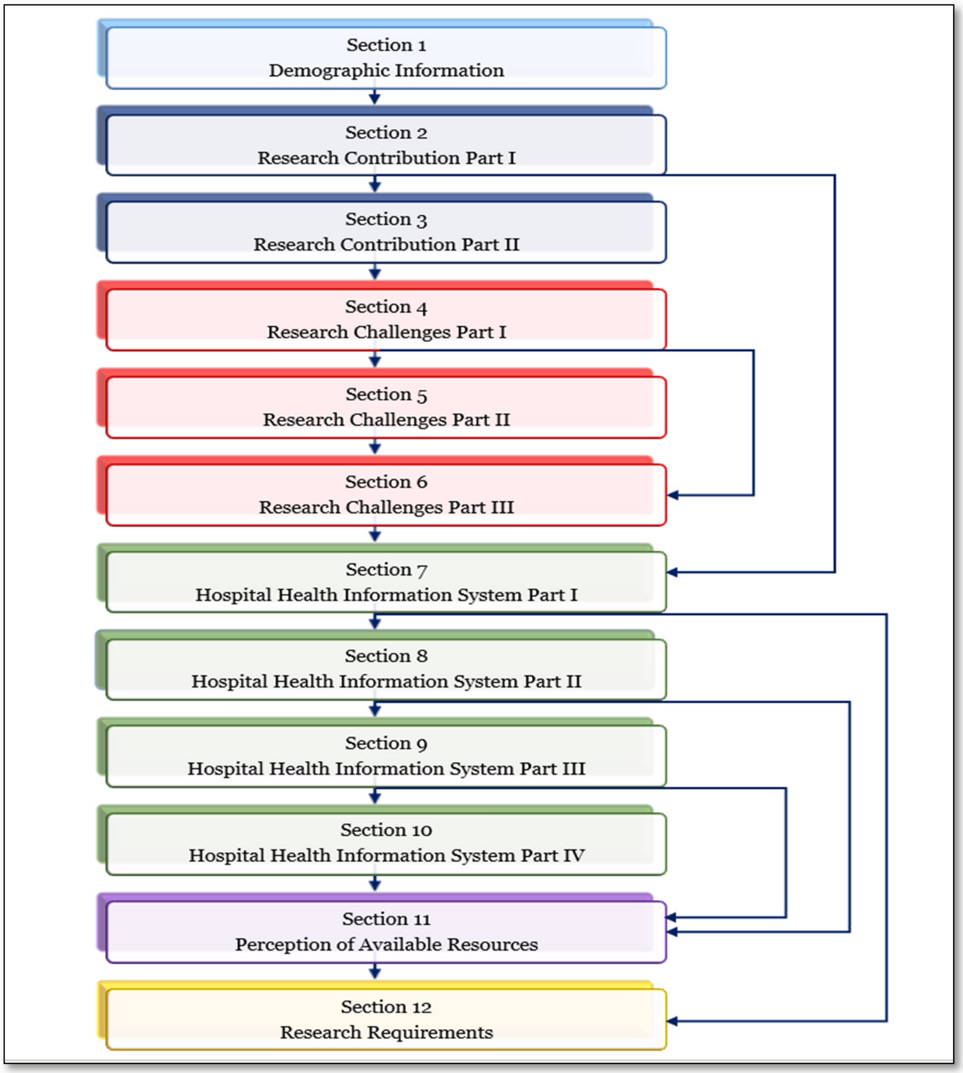

Section 1
Demographic Information
Section 2
Research Contribution Part I
Section 3
Research Contribution Part II
Section 4
Research Challenges Part I
Section 5
Research Challenges Part II
Section 6
Research Challenges Part III
Section 7
Hospital Health Information System Part I
Section 8
Hospital Health Information System Part II
Section 9
Hospital Health Information System Part III
Section 10
Hospital Health Information System Part IV
Section 11
Perception of Available Resources
Section 12
Research Requirements

Figure 3.1: Questionnaire structure

The first section had three items that gathered demographic information. The second section included a question to determine if the participant had performed any medical research during their professional career. This section redirected participants to the next section, according to their answers. The third section contained questions about their research contribution (the type of data, source and method of data collection). The next three sections investigated any research challenges that participants might have faced. The following four sections gathered information about the participants' hospital information system. Section 11 investigated participants' awareness of available resources that might help them with their research. Finally, the last section presented statements regarding research requirements and participants were asked to respond with the most appropriate answer.

3.2.2 Interviews

Participants and procedure

Of the eight hospitals contacted, only two responded to the interview request: King Abdulaziz University Hospital and King Faisal Specialist Hospital & Research Centre. The General Director of the two participating hospitals assigned IT personnel to participate in the interviews. These IT staff members were experts in their hospital information systems and had good knowledge of database management and hospital information systems.

The first interview was conducted at King Abdulaziz University Hospital. It was a face-to-face interview, involving pre-prepared questions (available in Appendix 3.2) and recording answers. This first interview lasted for 11 minutes and 50 seconds. The second interview consisted of the same set of pre-prepared questions in a Word document, which was emailed to the nominated IT expert at King Faisal Specialist Hospital & Research Centre. The response, with the answers, typed adjacent to each question in the same Word document, was received within five working days after the interview questions were emailed out.

3.2.3 Expert opinion

Expert opinion is considered to explore issues related to the secondary use of clinical data for research in broad areas of the healthcare sector including the genetic clinics and research centres for hereditary diseases which are mostly assigned to university hospitals. A formal discussion was had by phone, followed by an exchange of emails, with Dr Jumana Al-Aama, a leading researcher in the field of genetic disorders and founder of the medical genetic unit of King Abdulaziz University and the head of the Princess Al-Jawhara Centre of Excellence in Research of Hereditary Disorders (PACER-HD) at King Abdulaziz University Hospital, Jeddah. The phone call lasted 10 minutes to understand the system available and the issues that researchers may face while using the current system in their clinics.

3.3 Data analysis method

Survey data were analysed using SPSS statistical software Version 22 and descriptive statistics reported for Sections 1–11. For Section 12, a reliability test was conducted using SPSS on the 5-point Likert-type scale items (Q24–Q32 in Appendix 3.1), and the Cronbach’s α value was found to be 0.732 (items = 9); an α value of at least 0.7 indicates internal consistency [108]. Responses elicited through the two interviews with IT personnel are reported in summary.

3.4 Results

3.4.1 Survey

Summary data

One hundred responses (20%, n = 503) were received over the two months (from 1 June to 29 July 2017, inclusive) via the online survey. Data were received from healthcare professional respondents from all eight hospitals. Participants from university research hospitals contributed the majority of the data (44%, n = 100), followed by respondents from the Ministry of Health (non-teaching) hospitals (37%, n = 100). A summary of demographic information (Q1–3, Appendix 3.1) is shown in Table 3.1.

Table 3.1: Demographic information of survey respondents

Professional title	% (n=100)
Specialist	32
Consultant	22
Academic Consultant	22
Resident	16
Other	8
Hospital	% (n=100)
University hospital	44
Ministry of Health hospital or healthcare centre	37
Private healthcare facility	5
Other	14
Years of experience	% (n=100)
Less than 1 year	14
2–5 years	31
6–9 years	27
10 years and above	28

Research contributions

Table 3.2 shows the results of healthcare professionals' research contributions (Q4–7). About 78% (n = 100) of the respondents had conducted research at least once during their career. These researchers used clinical data more commonly than administrative and laboratory data.

The results also show that more than half of the researchers manually collected their data using paper-based medical records; however, 47.4% (n = 78) used electronic medical records.

Table 3.2: Research contributions of survey respondents

Did you perform any medical research at any time during your professional career as a physician?	% (n=100)
Yes	78
No	22
What type of data was required for your research? (multiple answers)	% of selecting an answer (n=78)
Administrative and demographic data	43.6
Clinical data (diagnosis, treatment)	38.5
Radiology (imaging)	23.1
Other	2.6
What was the source of your research data?	% (n=78)
Paper-based medical records	52.6
Electronic medical records	47.4
By what method did you collect your research data?	% (n=78)
Data manually entered into spreadsheets (Excel, SPSS, Google-Form, etc.)	61.5
Computer-generated spreadsheets (Excel, SPSS, Google-Form etc.)	38.5

Research challenges

More than two-thirds (71.8%, n = 78) of the health professionals who undertook research experienced difficulties and problems while managing their data. The majority had a problem finding the right datasets for their research, and the other issue was integrating data with different formats from multiple sources. Data collection, analysis and storage, were the main obstacles researchers faced in the data management process. The results for research challenges (Q8–12) are reported in Table 3.3.

Table 3.3: Perceived barriers to research by survey respondents

Did you have any problems managing the data for your research?	% (n=78)
Yes	71.8
No	28.2
Did you have any problems finding the right dataset for your research?	% (n=78)

Yes	73.1
No	26.9
Did you need to combine data from different resources?	% (n=78)
Yes	74.4
No	25.6
Did you face difficulties/ problems integrating different data formats?	% (n=58)
Yes	70.7
No	29.3
What type of data was required for your research? (multiple answers)	% of selecting an answer (n=78)
Data collection	61.5
Data analysis	52.6
Data storage	12.8
Not applicable	5.1

Awareness of hospital information systems

Almost two-thirds of the participants acknowledged the digital storage of patients' data in their hospitals, with almost 90% (n = 75) of them being able to electronically issue orders for their patients' daily care. The results show that 64% (n = 75) of the respondents recognised the availability of electronic medical records in their hospital systems; however, 28% (n = 75) were not sure if electronic medical records systems were installed in their hospital system. However, 85.4% (n = 48) actually use electronic medical records to record their patients' data. Of those who use electronic medical records, 63.4% (n = 41) follow the standard coding to assign codes to diagnoses and procedures, such as the International Classification of Diseases (ICD-9/ICD-10), whereas about 20% (n = 41) were unaware of whether any standards were used in their hospital system. The results for awareness of hospital information systems (Q13–20) are summarised in Table 3.4.

Table 3.4: Health professionals' awareness of their hospital information system

Does your hospital keep patients' data in digital form, i.e. (stored in computers)?	% (n=100)
Yes	65
No	25
Don't know	10
Can you electronically order medication for your in-patients?	% (n=75)
Yes	89.3
No	10.7

Can you electronically order medication for your out-patients?	% (n=75)
Yes	86.7
No	13.3
Can you electronically order tests, referrals and investigations, and organise appointments, etc., for your in-patients?	% (n=75)
Yes	90.7
No	9.3
Can you electronically order tests, referrals and investigations, and organise appointments, etc., for your out-patients?	% (n=75)
Yes	90.7
No	9.3
Has your hospital implemented electronic medical records to collect patients' data?	% (n=75)
Yes	64
No	8
Don't know	28
Do you use electronic medical records to record your patients' data?	% (n=48)
Yes	85.4
No	6.3
Don't know	8.3
Do you follow any standard coding system when using electronic medical records (i.e. the use of numeric codes that labels standard terminologies for clinical diagnosis, procedures, medications and allergies etc.), for example, the International Classification of Diseases ICD-9/ICD-10 coding standard?	% (n=41)
Yes	63.4
No	17.1
Don't know	19.5

Perception of available resources for research

Slightly less than half of the participants who had knowledge that their hospitals have databases for patients' data knew that there was local access to these databases. On the other hand, 36% (n = 75) could not confirm whether they are allowed to access the hospital database internally, whereas the remaining 17.3% (n = 75) were certain that local access to the databases was not available. The availability of online access to hospital databases was confirmed by 30.7% (n = 75) and rejected by 41.3% (n = 75), however 28% (n = 75) were not sure about this service. About half (48%, n = 75) had no knowledge of the provision of online access from their hospitals to other networks for medical research, whereas 34.7% (n = 75) said this service was not available. These results (Q21–23) are summarised in Table 3.5.

Table 3.5: Health professionals' perception of the available resources for research within their hospital environment

Does the hospital offer local access to their databases of patients' data for medical research?	% (n=75)
Yes	46.7
No	17.3
Don't know	36
Does the hospital provide online access to their database (patients' data)?	**% (n=75)**
Yes	30.7
No	41.3
Don't know	28
Does the hospital provide online access to any Collaboratory Distributed Research Networks or Clinical Data Repositories?	**% (n=75)**
Yes	17.3
No	34.7
Don't know	48

Agreed medical research requirements

Almost half of the participants strongly agreed, and 24% (n = 100) agreed with the statement "Data collected by hospital HISs are useful for medical research" whereas 17% (n = 100) were neutral. The majority of the respondents agreed that electronic medical records are often a primary source for both prospective and retrospective clinical studies, with the results for strongly agree, agree and neutral being 40% (n = 100), 37% (n = 100) and 22% (n = 100), respectively. The idea that is sharing patients' databases anonymously between hospitals results in better analysis and more accurate outcomes was supported by 85 respondents with 55% (n = 100) indicating they strongly agreed and 30% (n = 100) indicating they agreed, while only three respondents disagreed with the statement. Almost 90% (n = 100) of the participants agreed that the sustainable and incremental adoption of standards improves the quality of medical research. Also, 82 respondents emphasised the importance of standards-based electronic data storage for data integration for research purposes. The statement that data available from electronic medical records can be reused by researchers to answer different research questions was supported by 82% (n = 100) of the respondents. About 88% (n = 100) of the respondents believe the integration of hospital databases will enable the development and execution of medical research and allow for timely maintenance and the update of standards which will influence data quality and consistency. Eighty-four participants supported the statement "Virtual access to research databases increases the feasibility of medical research and resolves the issue of centralised hospital access". Table 3.6 summarises the extent of agreement or disagreement on medical research requirements (Q24–32).

Table 3.6: Agreed medical research requirements by survey respondents

Data collected by hospital Health Information Systems are useful for medical research.	% (n=100)
Strongly agree	49
Agree	24
Neutral	17
Disagree	5
Strongly disagree	5
Electronic medical records are often considered a primary source for prospective and retrospective clinical studies.	% (n=100)
Strongly agree	40
Agree	37
Neutral	22
Disagree	0
Strongly disagree	1
Sharing anonymous patient databases between hospitals will provide researchers with a large sample size for better analysis and accurate results.	% (n=100)
Strongly agree	55
Agree	30
Neutral	12
Disagree	2
Strongly disagree	1
The sustainable and incremental adoption of standards improves the quality of medical research by ensuring that analysis will be based on large volumes of relevant and up-to-date data.	% (n=100)
Strongly agree	46
Agree	43
Neutral	10
Disagree	1
Strongly disagree	0
The provision of data gathered from multiple EMRs into a single database will allow researchers to use the same datasets to answer different research questions.	% (n=100)
Strongly agree	48
Agree	34
Neutral	17
Disagree	1
Strongly disagree	0
Standards-based electronic data storage is fundamental to join and integrate multiple healthcare sources for medical research purposes.	% (n=100)

Strongly agree	54
Agree	28
Neutral	16
Disagree	2
Strongly disagree	0
The integration of multiple hospital databases will enable the formation and execution of medical research.	% (n=100)
Strongly agree	47
Agree	40
Neutral	11
Disagree	2
Strongly disagree	0
Integrated medical research databases will allow for timely maintenance and the update of standards to ensure the consistency and quality of data for medical research.	% (n=100)
Strongly agree	50
Agree	38
Neutral	11
Disagree	1
Strongly disagree	0
Virtual access to research databases increases the feasibility of medical research and resolves the issue of centralised hospital access.	% (n=100)
Strongly agree	51
Agree	33
Neutral	15
Disagree	1
Strongly disagree	0

3.4.2 Interviews

Although the health information systems used in Saudi hospitals differ, they follow the same standards of coding (ICD-9 and ICD-10) with the process completely moving to ICD-10 in the future. The hospitals covered in the interviews (Appendix 3.2) provide health information systems in all locations and departments with a strict policy of system use as it is compulsory to enter care information in the health information system. Local databases are available in all hospitals, and data can be extracted for researchers, but they have to request data that support their case from authorities, and this process requires approval from multiple channels. Online access to hospital data is very restricted and requires management approval. Therefore, the main obstacles that a researcher may face are data availability, ease of use and data quality (due to physicians being unsure as to the correct coding for a specific disease). Data centres are used to host servers for both clinical and administrative data; however, only at KFSHRC, a data warehouse is used to store and extract patients' data for research purposes. Also, the Health Level 7 standard protocol is used to integrate all the clinical care resources within KFSHRC. On the other hand, neither the hospital nor the research centre supports health information exchange with other Saudi hospitals. But the interviewee at King Abdulaziz University hospital said: "it is one of our future plans to have universal access, so all hospitals use the same database".

3.4.3 Expert opinion

The discussion and the written responses from well-informed sources (Dr Al-Aama and the assigned researchers) provided all the information and answers to the detailed enquiries about the current process of conducting research and the data availability and challenges within their clinics and research centre system. The overall communications summary was as follows: patients' information is mainly kept in paper-based records, so the data entry is done manually by researchers using traditional Excel spreadsheets to store the gathered data from paper-based documents (results and reports) or a hard copy of patients' information printed from the hospital system, and it is not possible to directly extract information related to a group of patients with specific common conditions from the hospital system. Therefore, the present system in the genetic clinic or a research centre lacks the basic capacity to handle a simple research question. Although the available system allows researchers a single selection for a specific condition, it does not provide even a simple query answer of more than one field as the patients' data is stored in a flat-file structure with no existing relationship.

3.5 Discussion

The literature on the Saudi context in Chapter 2 shows that there are few sources for obtaining clinical and genetic data for use in research. However, these sources are ineffective either due to the difficulty of accessing them or to the use of traditional methods of data collection, which makes them unfit for use in research. For instance, in a significant area such as genetic clinics and research centres, researchers depend on paper-based records for keeping patient data and use Excel sheets for managing their research data. Examining the area of medical research databases using a survey, interviews and expert opinion, six themes were identified on the use of health information systems for medical research in Saudi Arabia: difficult data collection, difficult data storage, lack of system interoperability, incorrect datasets, poor data analytics, and negative perception of the usefulness of systems. The results show data collection and data storage were two of the main obstacles healthcare professionals faced in managing the data required for their research.

Theme 1: Data collection difficulties

Data collection difficulties may be caused by access control and rudimentary collection tools. Specifically, difficulty in gaining access to local data in hospital databases due to restrictive administrative policies can prevent researchers from specifying the form of the extracted data where it is controlled by IT personnel. However, it should be noted that a balanced approach (a practical IT security and privacy policy) towards access control is part of good governance in data security and privacy.

Theme 2: Data storage difficulties

Data storage difficulties can arise due to the many different data sources and formats and a lack of comprehensive adoption of clinical standards that facilitate health data sharing and health information system communication. As a result, data silos are common because most hospitals keep their data in isolated storage, which in turn makes the integration process challenging. This can be further aggravated by the absence of linkage standards that support future interoperability between hospital systems, which is critical for data aggregation for research.

Theme 3: Lack of system interoperability

A lack of system interoperability leads to data unavailability, which is one of the major drawbacks of research. If hospitals do not support information exchange, their medical researchers will not be able to access the datasets of interest to investigate their research questions. Because of the issues outlined in the themes above, it was difficult for researchers to find valid and adequate data for research cases, because hospital data may not have covered all medical conditions and hospitals did not support data exchange. Invalid data can lead to incorrect datasets (theme 4), while inadequate or invalid data, or both, can lead to poor data analytics (theme 5).

In the end, it was concluded that themes 1, 2, 3, 4 and 5 collectively led to a negative perception of the usefulness of health information systems (theme 6) among healthcare professionals in medical research.

3.6 Recommendations

Although there are a number of significant technologies and emerging models that have contributed to data generation in healthcare information technologies and applications, many organisations continue to face difficulties with collecting, storing, analysing and retrieving health-related data [109]. Medical research depends heavily on data gathered from routinely collected electronic patient records, such as electronic health and medical records, which can be stored in hospital databases. Based on the results of this study and the six themes discovered, this research recommends the following courses of action for medical research and health information systems in Saudi Arabia. At the same time, these recommendations will necessarily face barriers to implementation, such as lack of consensus among stakeholders and limitations on funding and resources.

3.6.1 Recommendation 1: Improve data collection and storage (themes 1 and 2)

The proper collection, management, storage and use of health information plays a significant role in supporting research databases with reliable data that can improve the quality of healthcare services. Health information systems collect information on the daily activities from all individual departments, including patients' data, to be stored in operational databases which do not support medical research. Therefore, one solution to overcome this problem is to create a data warehouse by designing and scheduling the regular extraction, transfer and loading of patients' data from operational databases to research databases [110]. Legal and ethical issues about data warehouses need to be addressed according to the organisation culture, but this is outside the scope of this study. As health service providers look to the cloud environment as a key solution for e-health service delivery which is an ideal model for the growing demand for system integration, there is a need for a healthcare system as a service (HaaS) [111]. Cloud computing can potentially provide an ideal platform and powerful support to store and process clinical data for research. Therefore, the cloud environment can be efficient for information collection, sharing and establishing an integration platform that can combine patients' data from various hospitals. However, access to aggregated clinical data on the cloud needs extra precautions to maintain security and preserve privacy. The Saudi government needs to play an active role to fund cloud infrastructure and use legislation to protect health data security and privacy. It should be noted that data security and privacy is a crucial issue in medical research, but outside the scope of this study.

3.6.2 Recommendation 2: Address interoperability issues (theme 3)

Terminologies, vocabularies and classifications are important to specify common concepts and procedures in the medical and healthcare field, as pre-agreed specifications allow unrelated systems to integrate [112]. Digital patient care records that embrace standardised systems for capturing and storing clinical observation, such as Health Level 7 messages, can enable integration and data reuse for research purposes [111]. As Saudi health information systems are highly heterogeneous, there is a need for appropriate protocols and standards of health information exchange, such as Health Level 7, to ensure interoperability between multiple systems [112]. These standards would enhance the ability of the system to efficiently interpret and use exchanged information for further aggregation into larger pools to support robust analyses [111]. Therefore, collaborations between hospitals to adopt uniform standards to support interoperability and improve data exchange and integration process are necessary. The Saudi government needs to play an active role in mandating interoperable systems for medical research.

3.6.3 Recommendation 3: Overcome incorrect datasets and poor data analytics (themes 4 and 5)

Electronic health records contain all historical patient health information from previous visits and follow-ups, including links to radiology reports, imaging and laboratory results [113], [114], and electronic medical records can be considered as a data provenance for electronic health records [111]. Health records are a fundamental source of data for exploratory research, and they also support research studies in public health, outcomes research and epidemiology [115]. However, electronic health record data are not generated and maintained to address the needs of researchers, and issues such as incomplete, fragmented or erroneous records, which result from poor documentation due to physicians' workloads, and the lack of interoperability in electronic systems undermine their usefulness for research purposes [116]. This could be resolved by developing standard strategies that can deal with partially accurate and incomplete data. Theoretically, electronic health records could be shared across multiple hospitals [111], but to seamlessly connect systems, the adaptation of standardised code and data linking would be required. This would improve electronic health record use and facilitate the data collection process for research [117]. The Saudi government needs to lead to achieve consensus datasets and agreements among all stakeholders, so the datasets are valid, adequate and exchangeable or interoperable.

3.6.4 Recommendation 4: Overcome the negative perception of system usefulness (theme 6)

Once the above recommendations have been successfully implemented, the level of awareness of the importance of and confidence in using health information systems for medical research will increase, which in turn will reduce the negative perception of the usefulness of health information systems.

3.7 Pilot study findings

The pilot study revealed issues on the use of health information systems for medical research in Saudi Arabia. Data quality and integration issues in health information systems, such as electronic medical records systems, create major barriers to developing databases and data warehouses for research purposes. Poor data management methods for data collection and storage make it difficult for researchers to access the most appropriate datasets to investigate their research questions. This may also lead to invalid or incorrect datasets compromising the data analysis outcomes. Storing data in isolated databases within each hospital is one of the obstacles to the integration of different data sources, especially in the absence of a sufficient application of standards. Therefore, this study identified two challenges: (i) lack of proper methods for managing research data in terms of data collection and storage are major influencing factors for medical research productivity in Saudi Arabia; and (ii) lack of data integration and sharing resulting from data silos and electronic health records lacking interoperability is also a crucial barrier that hinders the availability of large datasets for research.

Narrow the scope to genetic clinics and research centres in Saudi Arabia

The pilot study showed that the problem of using healthcare data for conducting research is shared by all researchers from the various hospitals who participated in this study. However, the discussion with the expert in the genetic field showed the severity of the problem in genetic clinics, and research centres are very profound and need an urgent fix due to the crucial role of these institutions that lack proper systems for the diagnosis and discovery of genetic conditions as well as performing research. These centres are mostly affiliated with non-profit university hospitals and depend on funding and donations from individuals or charitable institutions. They provide free services, including genetic testing for residents from major cities and rural areas. These clinics have helped to reveal genetic disorders and generate curated data which can be a valuable source for researchers if the data are collected properly and stored efficiently. Sharing such data could advance research and enhance healthcare quality by providing accurate information on commonly encountered inherited disorders and lower the incidence of genetic diseases using preventive measures such as premarital and preimplantation testing [118].

Narrowing down the scope of the study to a specific area of care makes it feasible to deliver a reliable and consistent solution which is easy to implement and validate. Therefore, the research targets the problem in a significant area of genetic disorders due to the high prevalence of genetic conditions in Saudi Arabia, where patients are diagnosed and monitored through genetic clinics and research centres.

3.8 Summary

This chapter presented a preliminary study conducted using a mixture of a survey, interviews and expert opinion. Using multiple methods increased the capability to answer broader questions and cover the issue of health information system usage for medical research from different perspectives of healthcare professionals, IT specialists and expert knowledge. The analysis identified key issues such as data collection and storage difficulties as well as system incompatibilities. Recommendations to overcome these issues were presented, as well as the overall finding of this study. This study highlighted understanding of a real-world problem which exists on a large scale in most hospitals in the health sector. Therefore, it is feasible to focus on one specific care pathway to deliver an appropriate and practical solution. The outcomes of this study constitute a major contribution to the thesis methodology, together with the research gaps identified in the literature review in Chapter 2.

Next, Chapter 4 presents the research gaps identified from the literature in Chapter 2, defines the problem statement derived from this chapter, and outlines the overview of the proposed solution.

Chapter 4: Problem Definition and Solution Framework

4.0 Chapter overview

This chapter introduces the definition of the problem and presents the solution. Section 4.1 presents the purpose and goals of this chapter and provides links to previous chapters. Section 4.2 outlines the research gaps identified from the literature review and the pilot study. Section 4.3 presents the challenges identified by the pilot study. Section 4.4 states the problem statement and the initial investigative research questions. Section 4.5 demonstrates the analysis and evaluation of design methods for the ***What*** questions from the problem statement in the area of data management systems and an integration framework. Section 4.6 organises the solution framework for the thesis. Finally, Section 4.7 concludes the chapter and introduces the next chapter.

4.1 Introduction

This chapter presents the final step in the problem analysis stage of the thesis methodology in the real world and the first step in the design stage to prepare a theoretical framework of the solution in the abstract world (Figure 1.1, Figure 1.2). The outcomes from the literature review in Chapter 2 and the pilot study in Chapter 3 form the bases for the problem definition in this chapter. The purpose of this chapter is to summarise the gaps recognised in reviewing the literature in health information systems application and technology in Saudi hospitals and literature on performing medical research in the Saudi environment. It presents the problem statement which poses four major research questions (***What*** proper methods and ***How*** can a solution be designed). However, the two ***What*** questions are discussed according to the sub-questions which are derived to facilitate the design and implementation process of both the data management system and the integration framework, and to provide the steps taken and remaining steps in a complete solution overview.

4.2 Research gaps in the literature

A comprehensive literature review was conducted in the area of e-health application, health information system implementation, and health informatics development in Saudi hospitals, in addition to a pilot study to investigate the role of health information systems in medical research. The findings showed an area of the research lacking sufficient contribution from researchers in terms of presenting a problem and providing a viable, practical solution.

- Previous studies in health information systems in the Saudi context were limited to the implementation, effectiveness and impact of health information systems as well as identifying adoption barriers and defining successful models from other nations and presenting recommendations to overcome these obstacles.
- Few studies presented a theoretical framework using modern technologies to model a solution for some specific problems. However, the majority of the research studies are descriptive and lack of applied solutions.
- None of the studies has investigated the effectiveness of implementing health information systems based on health informatics frameworks in using healthcare data to leverage the care workflow, assist in decision making, data sharing and integration, and improve the research process and outcomes.
- These studies are not able to present a solution to solve the data management issue in healthcare settings, and they lack technical models to solve health information system issues such as integration mechanisms for a new system fusion to legacy systems.
- To the best of our knowledge, no existing studies describe, or present technical systems work within the health informatics framework to serve diagnosis data management, assist with diagnosis decisions, and act as a reliable source for research studies.
- Few studies discussed the issue of e-health adoption and integration theoretically, and there is no technical solution to demonstrate any physical or logical integration for health information systems across several organisations.

4.3 Challenges identified from the pilot study

A pilot study conducted using a mixed approach of a survey, interviews and expert opinion to investigate the extent of the health information system used in medical research in Saudi hospitals [119] identified two main challenges. Firstly, the lack of proper methods for managing research data in terms of data collection and storage is a major influencing factor in medical research productivity in Saudi Arabia. Secondly, the lack of data integration and sharing resulting from data silos and health information systems lacking interoperability is a crucial barrier that hinders the availability of large datasets for research.

The preliminary study was devoted to forming the problem and narrowing the scope of the study to focus on a case study of a specific care pathway. The genetic disorders diagnosis process is identified as a significant area in the medical domain worthy of further exploration and intervention.

4.4 Problem statement

The application of health information systems in Saudi hospitals varies where some hospitals lack computerised systems, and other hospitals use systems from different vendors [19]. Difficulties in implementation result in many issues in the Saudi health system, affecting healthcare professionals in their daily workflow as well as their research contribution due to the lack of quality data. The problem of clinical data provision for researchers in medical studies in Saudi Arabia exists at multiple levels. First, there is an inadequate implementation of electronic medical records, a lack of a standard format of data collection, and poor traditional data collection and management methods used for medical research [4]. Second, there is a lack of standards and integration profiles to enable information exchange among healthcare information systems, as the healthcare system in Saudi Arabia is delivered by multiple agencies, both governmental and private organisations, with different data protection and sharing policies [6].

Saudi Arabia is one of several countries that is affected by inherited diseases due to the high rate of consanguinity, which is a powerful factor shaping genetic disorders [120]. Saudi Arabia is keen to reduce the occurrence of genetic disorders through premarital and prenatal medical examinations, including genetic screening tests [121]. The Saudi authorities created an environment for research and training and established centres of excellence in different regions of the country [122]. Genetic studies conducted on the Saudi population have contributed significantly and effectively to the global effort to identify the common mutations of autosomal recessive genetic disorders in particular [123].

Genetic clinics and research centres provide free genetic diagnostic services, including genetic testing to Saudi citizens. These centres have helped to reveal genetic disorders and generate curated data which can be a valuable source for researchers if the data are collected properly and stored efficiently. Sharing such data could advance research and enhance healthcare quality by providing accurate information on commonly encountered inherited disorders and lower the incidence of genetic conditions using preventive measures such as premarital and preimplantation testing [118]. However, these centres lack adequate systems for managing and archiving patient data for research, making it difficult to integrate and build a national genetic diseases database [7].

Clinics and research centres lack an efficient data management system for care and research and lack an integration framework between centres which prevents a unified view for genetic studies and the generation of a Saudi specific genetic disorders database.

Developing a system based on a health informatics framework could help physicians and researchers better manage the diagnostic workflow, make informed decisions, and reuse the diagnosis data readily in medical research and public health. Also, using efficient technology to integrate multiple sources with these diagnostic data, such as using cost-effective new generation databases which have the ability to accommodate healthcare data heterogeneity, could provide large datasets for population-based research and establish a national genetic disorders database.

This research explores options for developing (i) a data management system for an individual genetic clinic and research centre considering its use in clinical care and research; and (ii) an integration framework for data from these clinics and centres, taking into consideration design factors such as applying health informatics frameworks, suiting a low-resource setting with limited ICT infrastructure and exploiting efficient technologies for cost-effective application. The problem statement poses four major research questions, and a full analysis of possible solution approaches is conducted to suggest an appropriate design approach. The four essential questions cover the ***What*** proper methods and ***How*** can a solution be designed for a data management system and an integration framework for these systems.

The fundamental research questions

Questions for developing a data management system and an integration framework:

1. ***What*** are the proper design methods for a data management system for a genetic clinic or a research centre considering its use in clinical care and research?
2. ***What*** is the appropriate design approach for integrating genetic data from genetic clinics and research centres in a low-resource environment and limited ICT infrastructure?
3. ***How*** can a data management system be designed for a genetic clinic and research centre considering a health informatics framework?
4. ***How*** can an integration framework be designed for aggregating genetic disorders data from multiple genetic clinics and research centres depending on efficient, cost-effective technologies?

4.5 Research questions (*What*)

The problem statement poses four fundamental research questions for the design and implementation of (i) a data management system for a single clinic or research centre; and (ii) an integration framework for aggregating data from multiple genetic clinic and research centre systems. The purpose of this section is to demonstrate the two ***What*** questions on the theoretical concepts, methods and approaches of the possible solution approaches from the literature. However, the ***How*** questions are demonstrated in the actual design of the proposed system in Chapters 7, 8 and 9. To facilitate the design solution process, each research question is subdivided into interrelated questions to serve the intended objective. Therefore, each research question is discussed individually according to its objectives to answer the question which evolved through critical analysis to the possible solution approaches in the literature and is then discussed in a comprehensive and detailed manner, showing the positive and negative aspects of these solutions. The discussion and analysis led to the optimal solution, which was weighted considering the factors stated in the problem statement such as health informatics frameworks application, suiting a low-resource setting with limited ICT infrastructure and using technology for efficient application.

Table 4.1: Research questions and objectives

PS.	Research Question	Sub-questions	Objectives
Data Management System (DMS)	1. ***What*** are the proper design methods for a data management system for a genetic clinic or a research centre considering its use in clinical care and research?	1. Why do they just use the HIS from their affiliated hospitals? 2. Why do they develop a clinical research database? 3. What is the proper data storage model for their database? 4. What are the efficient design methods for the system development as well as the database design? 5. What are the best evaluation methods for successful implementation considering health informatics requirements and user satisfaction?	1. Evaluate HIS use for research 2. Evaluate clinical research databases 3. Evaluate data storage models 4. Evaluate system design lifecycle methods 5. Evaluate HIS successful implementation models

Integration Framework	2. ***What*** is the appropriate design approach for integrating genetic data from genetic clinics and research centres in a low-resource environment and limited ICT infrastructure?	1. What is the approach for integration that fits integrating data in a low-resource environment with the purpose to develop a central database? 2. What is the cost-effective storage mechanism that is suitable for integrating heterogeneous data in limited ICT capabilities setting? 3. What is the best choice of data modelling techniques for better integration and fast query performance for research? 4. Why are ETL tools not a feasible option for moving data from source to destination, and how will this thesis overcome this obstacle?	1. Evaluate integration approaches 2. Evaluate data storage methods 3. Evaluate data modelling techniques 4. Evaluate the Extract, Transform, Load (ETL) process

4.5.1 Design of a data management system

The process of designing a data management system for use in healthcare and research requires paying attention to the health informatics principles; therefore, design, build and evaluation always start with a definition of a purpose for the system which affects its success or failure [26]. Informatics focuses on supporting healthcare systems, so it is a problem-driven endeavour. Thus, the system design process requires a full understanding of the nature of information and communication problems in the healthcare workflow. The problem arises from the health information systems application when it fails to satisfy health professionals because it has not been designed to meet the needs of the clinicians who will use it. The design of information and communication systems is not limited to software and hardware but must include the people who will use them [26].

4.5.1.1 Evaluate health information system use for research

Health information systems software tools such as electronic medical records for a single healthcare organisation are used for data collection, storage and management. However, the purpose of data collection was not for research, so the secondary use of data in research without thorough data cleansing and quality validation processes is known to be unreliable due to poor quality data issues of confounding, biases and missing variables [124]. The electronic health record, which can be implemented across multiple healthcare organisations, emerged with the initiative to improve the integration and availability of patient data [125]. Although the larger benefits are from implementing electronic health records as a tool for efficiency and standardisation, adequate implementation can be costly and time-consuming [126], which was the most significant contributing factor to the failure of widespread adoption [127]. Therefore, health information system methods of data collection and storage are not the best option for secondary use of the collected data for research, besides the well-known adoption and implementation challenges in Saudi Arabia [128].

Discussion and implications

The use of health information systems such as electronic health and medical records as a system for data collection, storage and management in any genetic clinic or research centre is not the best option, due to the costly setting and implementation challenges in Saudi Arabia, especially for limited budget small clinics and research centres that are based on charitable funding. Also, these systems do not support the information management cycle in these genetic clinics' daily activities of diagnosing conditions and do not support research.

The feasible solution for a single clinic and research centre is to design a novel system customised to serve management goals and manage healthcare activities as well as research taking into consideration user requirements, healthcare process requirements, and the environment infrastructure capabilities.

4.5.1.2 Evaluate clinical research databases

Health data aggregation can drive and support quality and safety improvement of healthcare services, public reporting, health services research, clinical research and public health [129]. Existing health information systems are not an available source for research due to their structure and method of storage as they store patient data in the form of reports and tables.

Researchers have to use a manual method of data collection and use traditional file formats for storing their research data [130]. Therefore, in clinical research studies, there is a need for a more efficient and effective way in terms of time for storage, retrieval and analysis of patient data. Clinical research databases serve a critical role in keeping patients' data in a structured format to be used in research; clinical data can be either collected manually or automatically transferred from health information systems. Clinical research databases tend to be specific to a disease, population, procedure, treatment or device [131]. The automatic transfer of patient data between patient care databases and clinical research databases can help reduce data duplication and increase consistency [132]. The structural organisation of the clinical research database is designed to support data retrieval and answer research questions using software tools for custom queries, reporting and statistical analysis.

Clinical data stores vs clinical data warehouses for medical research

Clinical data stores suit the daily routine of clinical practices for the patient care within each healthcare organisation. They contain disparate information across various departments and laboratories. To gain insights, distributed data stored in silos need to be integrated which is a difficult task due to the heterogeneity of the data sources that challenge the integration process which requires the development of a central interface for all systems and applications [133]. Data warehouses were always the best option for unifying scattered data in operational or transactional systems and provide a single view for useful, timely analysis for higher management and researchers. Thus, clinical data warehouses provide efficient storage and powerful analysis tools to support healthcare providers' decisions and answer researchers' queries. Although a clinical data warehouse can serve the purpose of data provided for research, it is difficult to build and requires many organisational resources for implementation and training purposes which means it is not a feasible solution for a small single clinic or research centre.

Discussion and implications

Data storage options such as clinical data stores or warehouses which serve different purposes for storing daily clinical practices and for reporting and analysis view respectively are not sufficient on their own to support the requirements of a genetic clinic or research centre for both research tasks and handling daily patient data management.

The best solution to this case is to develop a system with hybrid characteristics to support patient data management such as electronic data capture forms for patient data entry, update, delete and retrieve operations, as well as additional reporting and analysing features to support clinical decisions and allow physicians and researchers to gain more insights about their patients' common characteristics. This solution can be implemented as a web-based interface for multi-user support and a single data repository for analysis and reporting.

4.5.1.3 Evaluate data storage models

Entity attribute value model vs relational model for clinical research database

The relational model simplifies the representation of the clinical care process using modelling representation diagrams such as the entity-relationship model which graphically represent the conceptual schema which can be easily converted to the logical and the physical model. Relational models are easy to communicate and engage users and developers through the abstract representation (blueprint). Normalisation techniques allow for more flexibility and provide accurate query results. On the other hand, the entity attribute value model, which is the most widely adopted storage model in clinical systems, has a three-column fixed schema referred to as an entity, attribute and value which stores the primary key, the attribute name and the data value respectively. The entity attribute value model improves flexibility by allowing attributes to be added by just specifying their names in the attribute column [134]. However, the main model drawback is the restriction on a single value column which makes attribute-centred queries less efficient because of the large number of tables that contain many more rows than when using traditional relational databases [135].

Discussion and implications

Database design modelling approaches such as the entity attribute value model with a denormalised single value column attribute are not an efficient option to support patient-centred values with multiple attributes especially during the execution of a single value column attribute in large tables with numerous rows.

Thus, the relational model, with its conceptual, logical and physical modelling techniques, supports multivalued attributes which are common in most clinical descriptions. The model is easy to use and allows better communication and understanding for users during the development processes. Features such as normalisation provide flexibility and more accurate query results. This model will facilitate the execution of complex queries in an efficient time and reduce the time spent developing the system and allow more time for users to test and evaluate the system.

4.5.1.4 Evaluate system design lifecycle methods

The information system is a system which provides for data collection, storage and retrieval; it facilitates the process of transferring data into information and allows the management of both data and information. Hence, a complete information system consists of people, hardware, software, database(s), application programs and procedures [136]. The development lifecycle is a continuous process of creating, maintaining, enhancing and replacing an information system. Within the development framework, applications (program instructions) transform data into information for decision making. The database is a critical component in the information system, as the issues related to the design, implementation and management of the database are associated with the information system. Therefore, the performance of an information system is tied to the design and implementation of the database and the application, and to administrative procedures [136]. Successful information systems are developed within a framework known as the systems development lifecycle. Within the information, the system is a framework for database frequent evaluation and revision called the database lifecycle.

Systems development lifecycle

Systems development lifecycle is a general framework to track and understand the activities required to develop and maintain information systems. Within the framework, there are several methods to accomplish various tasks defined in the lifecycle such as entity-relationship modelling for a relational database, Unified Modelling Language (UML) which provides object-oriented tools, Rapid Application Development (RAD) iterative software development methodology, and Agile Software Development, which is a framework for developing software applications that divide the work into smaller sub-projects to obtain valuable outputs in shorter times and with better consistency. Although the development methodology may change, the basic framework within each methodology does not change [136]. The traditional systems development lifecycle is an iterative process divided into five phases: planning, analysis, detailed systems design, implementation and maintenance. The database lifecycle fits into and resembles the systems development lifecycle. The database lifecycle has six phases: initial database study, database design, implementation and loading, testing and evaluation, operation, and maintenance and evolution. Database design is not a linear process; instead, it is an iterative process that provides continuous feedback designed to keep track of previous steps. The database lifecycle starts by selecting a relevant approach to design the database from a set of business rules, processes and data that have been defined [137], such as *a data-driven approach*, *a process-driven approach* or a *parallel/blend* of both approaches where the focus remains on data, and business functions and rules throughout the development and implementation stages.

The development process involves several steps depending on the methodology followed, for instance, *the traditional method* (requirements analysis, data modelling, normalisation), *the Barker method* (strategy, analysis, design, build, documentation, transition, production), or other adapted design methods [137]. The conceptual part of the database design may undergo many variations based on two basic design philosophies: *bottom-up* (identifying the datasets, and then defining the data elements) versus *top-down* (identifying the data elements and then grouping them into datasets) and *centralised* (simple and small data components) versus decentralised (a considerable number of entities and complex relations) [136].

Discussion and implications

The structure of the data management system consists of two fundamental parts: the database which constitutes the backend of the system, and the application interface, which is the system frontend. Database design focuses on the use of the database architecture to store and manage data for end-users [136]. The principal purpose is to create reliable and useful data models that can effectively function as tools for communicating with potential system users and as blueprints for database developers. For the design process to produce a high-quality product for the customer, the basic steps of the database design process must be clear in a well-organised step-by-step guide to database design [137]. Before implementing design efforts, developers must understand the business processes and information requirements (rules, entities and attributes), and then convert the business attributes into a business model, which will later contribute to transforming the resulting business model into a database model using the design methodology [138].

The traditional systems development lifecycle for data management system design and implementation has five phases: planning, analysis, detailed systems design, implementation and maintenance. The entity-relationship modelling is considered as the relational database model is suggested for the storage of the data management system. The design focus is on both process requirements (diagnosis workflow) as well as the data requirements which are significant for the research use; therefore, the design approach of the database will comprise a parallel/blended approach of data-driven and process-driven approaches together. The Barker method is selected to organise the full cycle of the system development as well as the design of the relational database. Therefore, the Barker method guides both systems development and database lifecycle. The Barker method includes seven comprehensive design steps which exceed the traditional method by providing more steps that better organise the design effort for optimal functionality and performance. As for conceptual model development, the bottom-up approach is more productive in the centralised environment of genetic clinics with limited and identified data elements.

4.5.1.5 Evaluate health information system successful implementation models

During the implementation of the health information system, different organisational, technical and human barriers may occur at different levels of the implementation process [139]. Therefore, planning the implementation of any health information system in a healthcare environment regardless of the resource setting (low-resource or high-resource) must pay attention to critical factors to guarantee successful implementation and a high adoption rate. System quality, ease of use, responsiveness and security are essential factors influencing system implementation in a low-resource setting [11]. However, in a high-resource setting, additional regulatory or environmental factors influence the quality of the system and its uses such as leadership support for system design, development, deployment and ongoing support for user training [140]. Therefore, to investigate the successful implementation of health information systems and measure the adoption rate, evaluation methods have been developed or adopted in many studies based on frameworks structured around factors that positively influence the system implementation or adoption. The information system success model is used to define the effects of technological barriers on the user and determine the successful application of the information system based on three criteria: information quality, system quality and service quality [21]. However, the model focuses only on IT and system quality, and it is the only factor that determines the overall effect [141]. The technology acceptance model is also used to analyse user acceptance or rejection for a system based on two aspects, "perceived ease of use" and "perceived usefulness", to predict the behavioural intention of the user towards the system use [141]. Other models based on the idea of compatibility or "fitness" focused on the clinical workflow and the task required to be accomplished by the user using the IT, such as the task–technology–fit model which addresses only the fitness between user and technology and between task and technology, and the FITT framework (Fit between Individuals, Task and Technology) which is based on the fit between the attributes of users, technology and clinical tasks. In addition, the Human–Organisation–Technology Fit (HOT-fit) model is based on addressing the socio-technical components of information systems and the fit between them throughout the system development lifecycle [139].

Discussion and implications

Although there are few studies that have been conducted on the adoption of health information systems in developing countries, adopting models to assess the success and failure of implementation of systems highlighted real-world applications using case studies in developing countries and confirmed identified factors or defined new influencing barriers. Bawack and Kamdjoug [142]used a modified model of the technology acceptance model, the unified theory of acceptance and use of technology (UTAUT), to examine the behavioural intention of clinicians in Cameroon to accept and use health information systems. The authors confirmed the original unified theory was inadequate to describe the intention of the clinicians in developing countries, so they extended the model to fit the context. Also, it is found that the model performs better with the age of users as a single moderating factor [142].

Sustainability frameworks which are based on sustainability theories from literature focus on factors that influence the long-term usage of the system. Moucheraud et al. [143] developed a conceptual framework to explore the potential sustainability of electronic health information systems investments in Malawi, Zambia and Zimbabwe and to identify factors likely to contribute to the continuation of donor-supported initiatives after the original support had been modified or ended. Their finding suggests that although maintaining sustainability in a low-resource setting requires intensive efforts, long-term success will be reflected in the improvement of healthcare outcomes [143].

Afrizala et al. [144]developed a research framework to evaluate the implementation of the primary health care information system in a rural area in Indonesia. Their framework is based on two models: the HOT-fit model, which investigates the factors that influence management's decision; and the FITT model, which analyses socio-organisational factors that affect IT adoption in healthcare settings. Their results suggested greater interaction between human resources, infrastructure, organisational support and process factors is critical to effective adoption [144]. Multicriteria, user satisfaction analysis is used to measure user satisfaction and determine the strong and weak points of user satisfaction.

Kitsios et al. [145] implemented the MUSA (MUlticriteria Satisfaction Analysis) method to measure users' satisfaction with the e-appointment system of a Greek state hospital in Thessaloniki and obtained data to support decision-makers [145]. On the other hand, for a successful implementation, public health information systems need to be accommodated within an informatics framework model that supports health information exchange within the health community. The public health informatics framework developed by Gotham et al. [146] suits the dual-use nature of the information system; therefore, it was used to depict all phases of the public health emergency preparedness. Their finding emphasised the significant role of a well-established public health informatics framework for delivering an integrated information system infrastructure that enhances the effectiveness of the public health emergency preparedness [146].

In the successful development and implementation based on a service design approach of a widely used Portuguese electronic health record, a qualitative study performed after implementation showed that the electronic record was considered useful and easy to use, and these results are backed by widespread usage of the system. Teixeira et al. used the multilevel service design (MSD) approach to evaluate the design of the electronic health record for each stakeholder at three levels: service concept, service system and service encounter [147]. Multilevel service design accommodates the shared creative nature of customer experiences and enables experience integration from service concept design through to service system design and service encounter. The process involves four steps: studying the system user experience, designing the service concept, designing the service system, and designing the service encounter [148].

Good system design is important to ensure system functions work with their intended purpose. A well-implemented system also affects user satisfaction for sustainable adoption [9]. Poor user and organisational contribution in the process of the design and implementation mean not achieving the potential of health information systems. Lack of clinical engagement was one of the important reasons behind physician resistance which is one of the most crucial barriers to system adoption. Therefore, there are two evaluation methods to consolidate the database management system design and ensure successful implementation. Firstly, inspired by the successful implementation of the Portuguese electronic record, multilevel service design will be applied in this thesis as an evaluation method, and the four multilevel service design steps will be used to evaluate the lifecycle stages in the Barker method, as this model fosters user adoption at the implementation stage [147]. Secondly, an informatics evaluation framework that provides a heuristic for matching the stage of system development according to the system development lifecycle and the level of evaluation will be used [149]. Both methods of evaluation will prove how the iterative Barker methods used for the system development lifecycle have increased user experience with the system, and their constant feedback in each stage would improve the acceptance of the system in its finished form.

4.5.2 Design of an integration framework

Data sharing through databases is common practice for clinical research as data collected at multiple sites are integrated with disease-oriented database systems, since one location may not be able to collect sufficient data for analysis. Clinical institutions may also be limited in terms of research interests, so a common database can make the collected data available to researchers in a variety of locations [14]. The overall goal of data integration for the clinical research community is to be able to answer questions about aggregated data which can be very difficult if each individual data source must be accessed separately or sequentially. The objectives of data integration in the context of health information exchange, as stated by Nadkarni and Marenco [18] are:

- *Being able to look at the "big picture"*: Collaboration between institutions that perform identical or highly similar operations, but which are located far from each other, is necessary to be able to examine consolidated summaries of structurally identical data to compare their performance.
- *Identify common elements within different sources:* These can then be used as a foundation for interoperability between systems that use individual sources. Such an effort was made by the Unified Medical Language System of the National Library of Medicine (UMLS) which uses controlled vocabulary to achieve standardisation in certain biomedical areas.
- *Eliminate repeated efforts and errors as a result of non-communication systems*. Errors can exist in the same organisation if it uses multiple software packages from different vendors who can make communication difficult and lead to duplication and inconsistency of data throughout the organisation.

4.5.2.1 Evaluate integration approaches

Data integration from multiple types of data sources provides new knowledge using various datasets that cannot be gained from a single dataset [150]. Integration can be achieved using two broad strategies of physical data integration and logical data integration [18].

The physical data integration approach relies on the concept of copying the original data, which is reorganised and moved from one or more repositories depending on the scope, purpose and size of the data. The merged data is stored and managed by these new systems instead of the original source and is sorted in a single, queryable repository [18]. The physical integration approach architecture can be represented in the form of a data warehouse for a wider scope, and great analytical capabilities or a data mart for a small scope and special purpose focusing on one area may be used. The integration process starts with defining a global data model for the destination source. Then the selected data are migrated from the source to the destination using the extraction, transformation and load processes [18]. Ultimately, all integrated data are transformed into the structure required by the global model, which provides quick access and excellent response time for queries [151].

The logical data integration approach*,* also called virtual integration, uses conceptual schemes to bridge the representational heterogeneity of the databases and uses queries with the ability to collect and integrate data from distributed sources. The logical data integration architecture is based on data which are distributed in their original locations. In addition, intermediary software resides at a central location and uses a specific query protocol to communicate with the system that hosts the distributed data via the internet. The mediator software is the point of communication between the disrupted hosts and users and mediates their request for data. Data federation is used to represent the data in the logical integration strategy. To achieve logical data integration, a global schema is defined for use as a validation model for the user query. Next, the mediator uses the mapping information to identify the location of the desired elements for the requested query. Then the proper translation of the global query to the local database management system query language of the distributed sources will be performed by the mediator [18].

Discussion and implications

Logical data integration, also called virtual integration, is based on data which are distributed in their original locations. The mediator software which resides at a central location and uses a specific query protocol is the point of communication between the disrupted hosts and users and mediates their request for data. The global schema is used to validate the user query before the mediator uses the mapping information to identify the location of the desired elements for the requested query. The global query needs to be translated to the local database management systems query language of the distributed sources by the mediator. Logical integration approaches require persuasive communication and infrastructure capabilities among health organisations [18]. Currently, this approach does not achieve the objective of establishing an anonymised national database of genetic diseases data. It does not apply in Saudi Arabia, and the limitation in institutions' infrastructures and the currently adopted technologies do not provide an appropriate environment for the logical integration that requires highly advanced information and communication technology [98].

Physical integration: For integration solutions such as data warehouses that require a global schema, it is difficult to handle diverse sources with independent requirements or extremely complex schema to encompass all data requirements [53]. Therefore, researchers face the most significant barrier of insufficient technical infrastructure to support data collection manipulation and proper analytical tools. An integrated data repository with the traditional data warehouse structure and a predefined data model incorporated into the database schema constituted the major source for the data required for clinical and translational research, designed to integrate various aspects of patient care data [54]. In addition, integrating heterogeneous data with various formats and structures requires a flexible schema which is impossible while modelling conventional data warehouses [152]. Also, the design of integrated data repositories of medical data requires considering critical issues such as identity management, protection of confidentiality, convenient and flexible query, and semantic and ontology definitions at the schema level for data coming from different data sources with different naming and structural conflicts [55]. Data organisation and representation in integrated data repositories is very important for analysis, as they accommodate data from a range of various sources which may require a continuing extension. Therefore, the extension should be allowed without any major changes in the original schema to allow easy maintenance and data analysis [56]. Traditional integrated data repositories depend on a conventional data warehouse relational structure. Although the relational modelling offers many advantages for storage such as high consistency and availability, the performance decreases with data growth and faces scalability constraints as it is impossible to measure horizontally, and its vertical growth is limited [153].

The physical integration approach is the optimal solution for Saudi genetic clinics because the aim of the integration is to build a central database for genetic diseases. Therefore the design and implementation of an integrated data repository for genetic conditions prevalent in Saudi Arabia is the target goal of the integration framework.

4.5.2.2 Evaluate data storage methods for the centralised database

Relational databases (SQL)

Integrated data repositories are designed with initial objectives to enhance collaboration work and share best practices, to support a wide range of research in the medical field [154]. Most integrated data repositories are designed using standard data warehouse architecture with a predefined data model incorporated into the database schema. This traditional approach relies heavily on the data source to be modified to conform to the global schema [155]. The most critical step in the design of the repository is the data modelling process, which can be achieved by using modelling techniques such as entity–relational modelling and dimensional modelling. But the conceptual model in both modelling techniques should be carefully transformed to the logical design using one of the online analytical processing approaches Relational-OLAP, Multidimensional-OLAP and Hybrid-OLAP [156]. Generally, the process of creating the repository (data warehousing) is assumed to be a traditional domain of a relational database as it has been used to integrate different data sources with relational data storages, whereas non-relational databases remain unusual in data processing tasks. The design and implementation of the integrated data repositories which depends on the standard creation of the data warehouse require specific infrastructure and setups for ETL and OLAP processes as well as reporting tools; all these components are costly and time-consuming [157]. In addition, integrating heterogeneous data with various formats and structures requires a flexible schema which is impossible while modelling conventional data warehouses [152]. Modelling strict schema to handle and control the aggregated data into the new repository to maintain common data storage is one of the primary steps in relational database design. Even though this schema helps maintain a standardised format for the new storage (i.e. all data from multiple sources should be transformed and aggregated to comply with the target schema), it lacks the flexibility to be iterated easily and quickly when a new data source needs to be added. Therefore, integration solutions such as data warehouses that require a global schema may find it difficult to handle diverse sources with an independent requirement or extremely complex schema to encompass all data requirements and are unlikely to be changed to embrace future new sources [151]. Although relational database architecture provides numerous advantages such as high consistency and availability, its performance decreases while the data grow and it faces scalability constraints as it is impossible to scale horizontally and is limited to grow vertically.

Non-relational databases (NoSQL)

The non-relational concept of NoSQL (Not Only SQL) has crossed the boundaries of the relational model and provides alternative mechanisms to store and retrieve data by allowing data aggregation according to their usage (data access patterns). NoSQL databases (stores) rely on the concept that different data models based on structure and purpose have to be used in solving various problems [158].

- *Key-value stores,* such as Riak and Amazon’s Dynamo, where data is modelled as the simplest version of a pair of key and value, use the structure of “hash table” as the key is used to access the associated value and the key should be unique and well-ordered for better aggregation capabilities [159]. It is best for application in the area of highly volatile schema and requires storing lengthy values as the key-value store prefers high scalability over consistency which results in omitting features such as joins and aggregations and eliminating the possibility of executing complex queries [160].

- *Document stores,* such as MongoDB and CouchDB, use the same data model as key-value where a value associated with a key is the document's content coded in standard formats like XML (eXtensible Markup Language), JSON (JavaScript Object Notation), BSON (Binary JSON), etc. Document stores are best-suited for complex data structure with nested documents or lists and numerical values and applications that require the execution of dynamic queries [160]. It considers the document as a whole which can allow storing documents with different structure in the same set in another word schema-free structure [161]. Related documents or objects can be aggregated into a collection so they can be treated as a unit which is represented in a hierarchical structure. The automatic indexing feature for all a document's properties and the ability to group indexes by field names increases query efficiency [158].
- *Column family (Bigtable) stores*, such as Cassandra and HBase, have a "wide-column" structure similar to the relational database management system as the data are stored as a set of columns and rows. Still, unlike relational database management systems, there is no need for predefined schema [162]. Column family stores use row and column IDs as a lookup key, and this type of store is influenced by the original Google Bigtable paper [157]. Also, they adopt the sparse table data model where columns can be added to tables dynamically. The characteristics of storing related data in a column-oriented fashion enhance the performance of the aggregation function and support ad hoc queries [160]. However, column stores tend to store columns in a single table in the form of a key-value model [163]. The composite key-value data model can also be used to achieve a higher level of consistency when a record is distributed in multiple column families, so the composite keys consist of (record-key, column family-name and a unique column-name) [159]. The record-key or the row-key is used to define partitioning where the column-name can become dynamic to define the clustering. However, this structure results in limitations while scanning across hashed record-key values or trying to access partition columns [164].
- *Graph stores*, such as Neo4J and Infinite Graph, model the data as nodes and relationships to form a network. They consist of three data fields: nodes, relationships, and properties which take a form of key-value pair. It allows the indexing properties of nodes as well as relationships [158]. Most graph stores implement the W3C resource

description format RDF standard to create node IDs for each node, which makes it possible to integrate more than one dataset into a single graph store and execute queries using a SPARQL query language. They are suitable for discovering relationships between a large amount of data as well as finding patterns at a faster rate [157]. Graph databases can also capture ad hoc relationships due to the adoption of schema evolution [160].

Discussion and implications

Document store is the most popular, flexible, powerful and general NoSQL solution. It provides a dynamic schema to handle a vast amount of different data and accepts unknown formats or elements. Unlike the key-value and column family stores values that lack a formal structure and indexing capabilities and are not searchable, the document store is known for its strong support of data integration and analysis by features such as a rich data structure to provide a nested data structure and allow data aggregation according to their access patterns; and it supports dynamic queries by allowing secondary indexes to support users' direct queries and provide fine-grain access to any item within the document. The hierarchical structure of the document store enables fast extraction of any subsections of a large number of documents without the need to load the entire document into the RAM. This is different from the key-value store structure, where the whole document is stored in the value section and treated as a whole [157]. The most popular document store is MongoDB which provides solutions for data aggregation, handling semi-structured data, the flexible document store structure, and an API with an expressive, powerful query language that supports arbitrary field selection to handle ad-hoc queries. In addition, it provides scalability for a quickly growing data repository, where horizontal scalability is one of the NoSQL database features [160]. Furthermore, it provides an encrypted storage engine to meet Health Insurance Portability and Accountability Act requirements for identifiable (sensitive) patient data protection [165].

- The design structure contains collections which are equivalent to tables in the relational databases, and inside the collections, stored documents can be thought of as a row in a relational database management system but with a different representation of data and they do not have to follow the same structure or schema. Furthermore, more flexibility can be added by allowing new attributes to be created without the need to be defined or altering the existing structure of the documents [166].

- Automatic indexing for every property inside the new document when added allows querying over all elements and makes everything searchable and provides access to the exact location of any property using the document path to the leaf value [157].
- The structure of the document store, in general, is represented in the form of a tree structure, where the document trees have a root element(s) and underneath branches, sub-branches and values. Each branch contains a related path expression (document path) that explains how to navigate through the tree.
- The design focus is on the data access pattern, so the schema shifted into the application code. It is known as an implicit schema which is based on a set of assumptions about the data structure in the code for data management [166].
- The collection may contain all data required to be processed internally and displayed together. Embedded documents and arrays stored in a collection according to their access patterns reduce the need for joins between multiple collections providing for easy access and better-querying performance.

Document databases limitations

There are a few limitations in the general application of document stores, which does not affect the choice to use it as a backend solution for the integrated data repository system. It is difficult to perform an operation on a single level documents such as updating certain data within a document [166]. However, the design of the integrated data repository will focus on aggregated anonymised patients' data with common genetic diseases, so there is no need for automatic cross-document operations as all documents are written at once to the database and are ready for the read operation. Updated patient datasets at the source will also be uploaded in a single operation with no updates allowed at the integrated data repository. In volatile structure situations where the aggregate design is subject to frequent changes, the query process will be a difficult task [166], but the aggregated design structure which is based on a patient-centric structure for the integrated data repository will remain constant.

According to the previous discussion, the requirements of the proposed system would be compatible with the document databases, and the NoSQL document store is the best choice for the system design and the integrated data repository requirements such as:

Easy to combine data with different structures and formats from various sources using a schema-less document model.

Allow a unified view of the data (i.e. provide data insights) according to the data access patterns, in this case, the joining or grouping of the data that will be read together.

Support researcher needs to answer their research questions using ad hoc (dynamic) queries based on multiple fields.

Provide automatic multiple indexes, especially in a flexible query interface implementation in MongoDB.

Accommodate future data growth depending on the NoSQL scaling capabilities and powerful performance to handle big data characteristics.

Other features such as being less expensive to manage and more flexible in deployments make the NoSQL database attractive for any integrated data repository developer.

4.5.2.3 Evaluate data modelling techniques

Data modelling is an abstract process used to demonstrate our understanding of the world of problems and provide a high-level representation of data objects, their relationships, and the rules that govern operation on objects before delving into the details of the intended system [167]. The goal of the data model is to provide the basic concepts and notations that allow database designers and users to communicate easily and accurately and understand organisational data [168]. In general, the term data modelling is tied to the design of the schema or the logical model. This is true in relational database design where the schemas must be defined in advance before writing the data to the database. However, in NoSQL databases, it is not necessary to fully design the data model before inserting the data [169]. While modelling in the relational database is based on data structure, NoSQL database models rely heavily on data access patterns (queries to be executed on data). Although the NoSQL databases are schema-less, the significance of the data model to demonstrate and understand the storage is certain [160]. Together the logical model and queries capabilities determine system usage, where the physical data model is accountable for the system performance [159].

Design in NoSQL databases mostly focuses on application performance rather than actual business models and rules for integration and standardisation, which results in issues during the implementation of the databases. Instead, NoSQL developers skip the data modelling stage (i.e. schema-less) to creating the data storage and move the data instantly directly. Although it is the fastest way to start, it usually causes problems in both data storage and query performance. Applying a certain level of the schema to all or some of the stored data increases the quality of the stored data and ensures its purity to a certain level [168]. Therefore, many NoSQL database stores offer semi-structured models and list-like data types [159].

Discussion and implications

One of the main characteristics of NoSQL databases is that they are schema-free, so in document databases, there are no restrictions on the structure of the stored document. The data can be stored and managed without the need to develop data models such as the entity-relationship model. However, the design of NoSQL document databases supports direct aggregation of nested data structures [168]. Document databases such as MongoDB have a flexible schema which is written and applied by the application developers. When designing the data models, attention should be paid to the data access patterns, i.e. the application usage of the data including queries, updates and data management as well as the nested structure of the data itself [170]. Modelling techniques such as aggregation may be used to convert the logical model to the physical one. For modelling relationships (i.e. embedding data or referencing) documents may be applied according to the data access patterns or queries executions. Embedding data is merging data using hierarchy (parent-child relationship) in the case of one-to-one or one-to-many. However, referencing denotes a link created in one collection to refer to another collection similar to the foreign key function in the relational database management system [169].

Data modelling concepts for document-oriented databases

In the NoSQL document store databases, the appropriate adoption of effective modelling techniques of relationship styles such as embedding and referencing play a significant role in data modelling. There are many factors (such as relationships cardinalities notations, i.e. one-to-one, one-to-many and many-to-many relationships) to be considered when choosing between referencing and embedding a document [171]. For example, MongoDB data modelling allows both normalised and denormalised data models through the application of supporting tools to manage relationships by using references among data in different collections and embedded documents [172]. The decision to embed or to use references is a critical factor that will influence data model design and consequently affect application performance and database capacity [170].

Embedded data models, also called “denormalised” models, allow applications to store related data in a single structure or document (database record). MongoDB allows including documents within another document as a field or as a matrix. Applications may then need to issue fewer queries to accomplish joint operations. The embedded data models can be considered when there are “include” relationships between entities such as a one-to-one (customer and address). Here the pattern should be considered, in other words, how the data will be accessed or queried. For example, if the goal was to display entity data in the context of the other entity and all data from both entities are frequently displayed together, then the embedded data model is better than the normalised data model with references. While only one query retrieves the entire information through embedding, more than one query is needed in case of Referencing. Similarly, in one-to-many relationships (customer has more than one address) embedding the address data into the customer document is better than using a customer_id as a reference in the address documents. Overall, regarding the data access patterns, embedded data models provide optimised read operations, with quick and easy access to retrieve or update relevant data in a single atomic database process.

This modelling technique allows defining relationships between data within a collection. This approach uses references or links to manage a relationship between two entities, i.e. including the field of a document (document's_id) into another document as a reference. Then whenever the embedded data need to be accessed, a second query is issued to resolve the referenced fields [172]. Referencing can be an option when data duplication resulting from embedding reduces the read performance, and also to represent more complex relationships (many-to-many). Moreover, it can model large hierarchical datasets using a tree structure by storing a reference to parent nodes or child node [170].

Other operational factors should be considered when designing the data model, which may affect the performance of the application in terms of reading and write operations. Therefore, for more efficient queries, productive insertion and update operations, or more effective activity distribution among a shared cluster, details of features such as indexes, horizontal scalability, document growth and data lifecycle management should be kept in mind during model design [170].

Data modelling and preparing a logical model before the actual physical design is a key step in the development of a successful implementation of NoSQL databases. But using relevant modelling techniques that produce a smooth transition between the logical model and physical model is equally important. Therefore, based on the data access patterns or queries executions, this study uses modelling techniques for data aggregation using embedding and referencing methods to convert the logical model to the physical schema.

4.5.2.4 Evaluate the extract, transform, load (ETL) process

Extract, transform and load (ETL) means to take operational data from the transaction systems and manipulate it and then load it to a separate database data warehouse for reporting and analysis [173]. The term ETL is used to denote the full ability to transfer data of all sizes in a fast and reliable manner from one place to another in real-time or as a batch scheduled, and this can be achieved through an integrated set of tools from one vendor or a combination of tools from multiple vendors [173]. The ETL process is usually done at the staging area or transformation engine where all the complex mapping process takes place before moving data from sources to the destination. The design and implementation of the integrated data repository which depends on the standard creation of the data warehouse require specific infrastructure and setups for extract, transform and load processes and online analytical processing processes as well as reporting tools; all these components are costly, time-consuming, difficult to build and require extensive organisational resources for implementation and training purposes [157], [133]. The development of ETL and customising its use is very expensive, and maintenance is also a big challenge; therefore, implementing a system based on ETL is not feasible for low-resource small-and medium-sized organisations.

Discussion and implications

Based on the above discussions and suggestions, the source database management system will be developed using a SQL relational model, and the target integrated data repository storage will be designed on top of the NoSQL document database. Therefore, the data should go through pre-processing before residing in the integrated data repository. The best way to aggregate the component of the proposed system and design the integrated data repository is by extracting data from the sources using customised methods for data extraction and transformation to the NoSQL document-oriented database to be loaded according to the data modelling technique, and the schema described based on the query patterns.

The design of the integrated data repository incorporates the ETL process, which occurs at the source database management system in a genetic clinic. Novel mapping methods for the ETL will be customised to extract the required data from the source applying certain criteria, then transform the data format from SQL to NoSQL, then finally apply the method to load the transformed data to the integrated data repository.

4.6 Solution overview

The thesis solution framework is part of the thesis methodology presented in Figure 1.2 which is structured according to the thesis approach in Figure 1.1 to answer research questions in each stage of the thesis methodology, as illustrated in Table 4.2. Initial steps towards problem identification asked preliminary question 1 about health informatics in the Saudi healthcare systems, i.e. the role of health information technology in improving healthcare services to determine the adoption rate of health information systems and identify possible barriers to the successful implementation in health organisations. The existing literature was surveyed to understand the environment and any difficulties and to clarify the problem domain. Next question 2 aimed to direct the research focus and scope to narrow down to a specific area of the problem domain to be able to provide a rational solution. A pilot study was conducted to investigate the challenges related to the use of health information systems for medical research from the perspective of healthcare professionals and IT staff who work in these systems. A combination of three methods (surveys, interviews and expert opinion) was adopted to understand the nature of the problem in depth. The finding of the study provided a thorough understanding of the problem to demonstrate a clear problem statement which was the source for the remaining research questions.

Based on the problem statement, research question 3 and question 4 were introduced to determine the development methods and concept by critically evaluating previous literature for possible solutions. Within these major questions, several investigative internal questions are applied to allow critical evaluation of alternative options of methods for the design from literature. Following this stage, an overview of the proposed solution is clearly identified and presented in the solution architectural framework in Chapter 6. The actual design of the data management system, as well as the integration framework, is the answer to research question 5 and question 6, presented in Chapters 7, 8 and 9.

The research is motivated by the objectives to fill the research gap identified from the literature and to allow the study to be applied in a genetic diagnosis setting to meet the imperative need for an efficient data management system in genetic clinics and research centres in Saudi Arabia to improve their daily workflow in patient care and improve their research productivity and contribution to the local and global research communities. The thesis aims for each clinic and research centre to acquire a system that facilitates the process of collecting and storing patients' data to be effectively used in clinical research. It also promotes research data exchange between centres and allows data integration in a unified central repository to provide larger datasets for accurate analysis and reliable outcomes. Thus, it sets the foundation for establishing a national database of genetic disorders.

The thesis proposes a twofold solution:

- the design and implementation of a novel genetic disorders diagnosis data management system called G3DMS

- the design and implementation of a NoSQL-based integrated data repository for genetic disorders data called GENE2D.

Table 4.2: Solution overview

Research Question	Research Method	Purpose	Objective	Chapter
Q1. What is the current state of e-health and health informatics in Saudi Arabia?	Literature review (Qualitative)	Define problem domain	1. Investigate ICT application in Saudi healthcare systems 2. Examine HIT status in Saudi healthcare 3. Identify possible issues with HIS implementation in the Saudi healthcare setting 4. Summarise findings and highlight the problem domain	Chapter 2
Q2. What are the challenges that a Saudi physician faces regarding the use of information from HIS in medical research?	Pilot study Mixed approach (questionnaire, interviews, expert opinion)	Narrow the scope of the study	1. Investigate the current state of medical research in Saudi Arabia 2. Identify possible issues that hinder the improvement of medical research and the role of hospital databases in supporting the research 3. Identify possible solutions to enhance the role of HISs in medical research in Saudi Arabia 4. Select a specific area in the healthcare domain	Chapter 3
Problem Definition				Chapter 4
Q3. What are the proper design methods for a data management system for a genetic clinic or a research	Critical review, analysis, comparison, and evaluation	Evaluate design methods to answer research question 5	1. Evaluate HIS use for research 2. Evaluate clinical research databases 3. Evaluate data storage models	Chapter 4

centre considering its use in clinical care and research?			4. Evaluate system design lifecycle methods 5. Evaluate HIS successful implementation models	
Q4. What is the appropriate design approach for integrating genetic data from genetic clinics and research centres in a low-resource environment and limited ICT infrastructure?	Critical review, analysis, comparison, and evaluation	Evaluate design approaches to answer research question 6	1. Evaluate integration approaches 2. Evaluate data storage methods 3. Evaluate data modelling techniques 4. Evaluate the Extract, Transform, Load process	Chapter 4
Theoretical Considerations				Chapter 5
Architectural Framework for the solution				Chapter 6
Q5. How can a data management system designed and implemented for a genetic clinic and research centre considering health informatics framework?	Barker's method for SDLC & DBLC MSD method for evaluating SDLC (Qualitative) Informatics Evaluation Framework for evaluating the SDLC (Qualitative)	The design and implementation of a Genetic Disorders Diagnosis Data Management System (G3DMS)	1. Support the diagnosis workflow 2. Assist the physician when making the diagnosis decision 3. Allow appending legacy data to the new system 4. Allow reuse of diagnosis data in research 5. Provide a valuable source for healthcare data integration and sharing	Chapters 7 & 8
Q6. How can an integration framework be designed and implemented for aggregating genetic disorders data from	Physical integration approach (NoSQL document database/ MongoDB)	The design and implementation of a NoSQL-based integrated	1. Integrate multiple sources of genetic disorders data from genetic clinics and research centres	Chapter 9

multiple genetic clinics and research centres depending on efficient, cost-effective technologies?		data repository for Genetic Disorders Data (GENE2D)	2. Provide an easy-to-use query interface for researchers to conduct their studies on large datasets 3. Provide potential to help grow and develop a national genetic disorders database in Saudi Arabia 4. Contribute to the national public health informatics	

4.7 Summary

This chapter pursued the problem formation to formulate a clear and precise statement. The research gaps and findings from the pilot study helped in shaping the problem and developing the thesis problem statement. The lack of a data management system in genetic clinics and research centres affects the performance of healthcare professionals in the diagnosis process as well as their research productivity. Also, a lack of data integration and sharing among these centres prevents the establishment of a centralised genetic database, which affects national level genetic studies and public health outcomes. Four major questions are derived from the problem statement to help formulate the solution, and the What questions are further decomposed into sub-questions then discussed and evaluated to nominate the best technology, methods and approaches for the proposed design. The complete design methods and approaches to the solution are presented in Table 4.2. Finally, the thesis solution to the problem is revealed.

Next, Chapter 5 provides substantial evidence of the fundamental knowledge required for developing data management systems and their databases with a focus on the theoretical basis for building data management systems and databases.

Chapter 5: Clinical Data Management for Research Purposes: Theoretical Considerations

5.0 Chapter overview

This chapter provides the theoretical foundation and the understanding of theories and concepts relevant to the design and development of a data management system and an integration framework and the promising technologies for their deployments. Section 5.1 presents the purpose and goals of this chapter and provides links to previous chapters. Section 5.2 presents contemporary database design issues. Section 5.3 explains clinical data management issues related to using in research such as data collection, pre-processing, storage and analysis. Section 5.4 describes the clinical research database types and usage, while Section 5.5 details the structural designs for clinical databases and their design challenges. Section 5.6 presents effective technologies to improve clinical research database integration. Finally, Section 5.7 summarises the chapter and presents the next chapter.

5.1 Introduction

This chapter supports the theoretical framework of the solution in the abstract world in the design methodology in Figure 1.2 and presents a solid foundation of theoretical understanding regarding the proposed solution and helps to answer the ***How*** questions presented in Chapter 4. The answers to ***What*** questions yielded the overall solution framework. At the same time, this chapter aims to consolidate the architectural framework of the proposed solution by providing a comprehensive view and detailed structure of the discussed solution options in Chapter 4. Databases provide a single comprehensive view suitable for analysis and relevant information for a variety of organisational purposes. In healthcare, databases serve an additional critical role in several areas, including patient care, administration, research, financial and billing. The shared use of data promotes information consistency for research and decision making and reduces the duplication of data, and the time and effort required for data collection. Data integration (physically or logically) from a system with modern database management systems that are compatible with new communication protocols increases the ability of researchers and clinical decision-makers to run queries rapidly on large datasets. However, if the existing data residing in legacy systems or non-database sources, and the data format does not support multi-user queries such as spreadsheets or statistical packages, the integration process is impossible or prohibitively expensive. This chapter explores some theoretical considerations for the design and implementation issues related to clinical data management for research and clinical research databases to achieve the intended system integration, specification and user satisfaction.

Preliminary

Technological advances in data acquisition, processing, storage and management, especially in biomedical domains, have generated massive quantities of data which differ in type and structure according to their sources and collection methods. For maximum benefit, these data should be effectively collected, organised, integrated and stored in a shareable and accessible way. Usually, data in clinical research databases are collected either as part of the patient care process or extracted from patient medical records. Data are collected and stored in a database organised and operated by a database management system to facilitate analysis using specialised software tools from which conclusions can be drawn to influence future decisions [132]. For a database to serve multiple purposes, the content and description of the database should be comprehensively covered [14]. Numerous integration efforts have been made, such as using approaches to integrate terminologies, ontologies and schema matching [174]. The process of data integration requires combining scientific methods and specifications which need to be stored in a database. Interoperability provides more options for integration. Before exploring the theoretical considerations of designing a clinical research database in Sections 5.4, the following section presents contemporary database design issues.

5.2 Contemporary database design issues

Database design focuses on how the database structure will be used to store and manage end-user data [136]. The primary objective is to create good and useful data models that can function effectively as tools of communication with the user community and as database blueprints for database designers. For the design process to produce a quality product for the customer, it should clearly define the basic steps of the database design process and have a structured plan (design methodology) to provide a step-by-step guide for database design [137]. First, the business processes and information requirements (the entities, data and rules) should be understood before proceeding with the design. Next, the attributes of the business are converted into a business model. The resulting business model can then be converted into a database model using the design methodology.

5.2.1 Data modelling

Data modelling is a design discipline with the task of analysing the business requirements followed by design according to the requirements analysis outcomes. Data modelling, the first step in designing a database, refer to the representation of the data required to support a process or a set of processes [138]. Data modelling is an essential part of the design process and the development of a data system. Data modelling provides techniques for describing the real-world information requirements in an understandable manner to the users and supports designers to implement the information requirements into a physical database system. Data modelling is an iterative progressive process which starts with understanding the problem domain by collecting and analysing details about data elements and their suitability for supporting the business processes. The next step is ensuring that the results of the requirements definition are fully implemented as data contained in the database. Based on the requirements analysis, a proper database system will be selected. The final data model is actually a blueprint that contains all the instructions to build a database that meets all end-user requirements [136]. The data model can be fitted at any time during the database design lifecycle. Therefore, it can be produced before, after or in parallel/blended to the process model. For instance, for *process-driven approaches,* the focus is mainly on the process model; the data modelling process starts by identifying all the processes and their required data. Then the data model is designed to support the specific data requirements of a particular process. However, for *data-driven approaches,* where the data model is developed before the detailed process model, which promotes the reusability of data, a consistent set of definitions is established for data and language to classify the data. In practice, it is impossible to develop a data model without investigating the processes or developing a process model without considering the data. Therefore, *parallel/blended approaches* are the ideal choice for dealing with the interdependency of data and process modelling [138].

Data modelling approaches

Traditionally, the file system method of organising and managing data has proven superior over the manual archiving of files to keep important data. This method has many challenges, such as data retrieval and data modifications that require extensive programming in addition to security risks. Different models have evolved in terms of better data management to overcome the shortcomings of the file system [136]. Several standard techniques and notations exist for creating a data model. Every modelling method has a set of symbols to represent some aspect of the real world, and a set of rules and procedures for using the symbols as the information requirements should be represented in a clear natural language through understandable data examples and intuitive diagrams. Therefore, data modelling approaches such as entity-relationship data modelling, fact-oriented data modelling and object-oriented data modelling can be adapted to satisfy the data modelling objectives [175].

The hierarchical model: This model was developed in 1960 to manage complex data for projects such as the Apollo space missions. The hierarchical structure contains levels or segments that represent a set of parent-child one-to-many relationships between a parent and its children. The hierarchical model delivers various advantages over the file system model, and many of its features laid the foundation for the current models. However, the model is complex to implement and manage, and it lacks structural independence between conceptual and implementation models [136].

The network model: The network model was developed to represent complex data more effectively than the hierarchical model to improve database performance and to impose a database standard. The network model allows a record to have more than one parent. A relationship is called a set, and each set is composed of at least two record types: an owner record and a member record with a one-to-many relationship. The Database Task Group was created at the Conference on Data Systems Languages (CODASYL) to define standard specifications for the database environment creation and data manipulation. The task group report included standards for three database components: (i) the schema, the conceptual organisation of the whole database as viewed by the database administrator; (ii) the subschema, which defines the portion of the database as seen by the application programs; and (iii) the data management language that determines the environment where the data can be managed. Three language components were specified for the three database components: a schema data definition language, a subschema data definition language, and a data manipulation language to work with the data in the database [136].

Entity-relationship modelling: The most popular and widely used approach was first presented by Peter Chen in 1976 [176]. This method recognises the information requirements as a set of entities and attributes participating in relationships. Entity-relationship modelling conceptualises the information requirements separate from any database software and hardware considerations. Therefore, this method suits conceptual modelling; however, it lacks a clear indication of the constraints in the relationships [136].

Fact-oriented modelling: This approach presents information requirements in terms of objects playing roles where the role is the part played by one object in a relationship. The object-role modelling technique is an example of this approach. It represents all the facts in terms of entities or values and allows for relationships with many roles [175]. In object-role modelling, ellipses represent object classes, which are either entity classes (sets of entity instances) or domains (sets of attribute values). The same symbol (multi-compartment box) is used to represent the relationship between object classes as well as to model the attributes of an entity class. The lack of a distinction between entity classes and attributes results in many shapes and a very complex representation which is difficult for the stakeholder to understand compared to the entity-relationship or the Unified Modelling Language (UML) representation of the same model [138].

Object-oriented modelling: This approach, which is primarily used for designing code of object-oriented programs, can be adapted for conceptual modelling and database design. The elements in the object-oriented model are represented as objects which have assigned properties that can be modified or inherited from other objects [136]. The UML is the most popular and widely used object modelling technique for both conventional and object models. Although UML is an object modelling technique, not a data modelling technique, modelling object classes is similar to entity modelling. For this reason, it is considered to be a data modelling technique. The UML models object classes using associations instead of relationships, and they describe the behaviour of each object. The data models are represented by class diagrams which have a greater ability than the entity-relationship diagram in capturing a variety of data structures and rules [138]. The UML is often seen as a design tool for object-oriented databases, and it can also be used as a design tool if the target database management system is a relational database [177].

Other modelling trends: Early attempts were based on the hierarchical and the network models, where data are viewed as hierarchical segments in parent-child relationships in the hierarchical model, and in the network model, data nodes are arranged as a network. However, in the relational data model, data are presented in the form of two-dimensional tables; hence it faces limitations in representing unstructured data types such as images, sounds and spatial elements. Post-relational data modelling approaches are used to overcome the unstructured data requirements. For instance, *object-oriented databases* acquire additional features to support complex objects, encapsulation of process and data, and user-defined data types. *Deductive databases* are an effective approach to represent and manage complex data derived from a series of recursions. *Spatial databases* efficiently manage spatial data such as maps of lands and two- and three-dimensional designs such as town plans [175].

5.2.2 Process modelling

Data modelling is concerned with representing data elements in a consistent manner through the creation of conceptual, external, logical and physical models. On the other hand, other aspects, such as business processes that use the data elements, must be modelled. There are many ways to represent these processes using data flow diagrams, process flow diagrams and function trees as well as UML diagrams such as use-case diagrams, activity diagrams, sequential diagrams, collaboration diagrams and statechart diagrams [175].

5.2.3 Relational databases

The relational database model is the most common database model and is the basis for modern database management systems due to its capability to manage a large amount of data, its performance and its reliability. The relational database model depends on the concepts of "***attribute***" which refers to the name of a set of data that represent the same thing, "***domain***" which is the smallest unit in the database representing an individual value, "***relation***" which refers to a set of related attributes as defined by the user, "***tuple***" which is a set of related data within a relation, and "***primary key***" which is an attribute to show that the domain value is unique among any tuple of the relation [177]. The primary unit of the relational database is a table which consists of rows "records" (related to an individual record in the table) and columns "fields" (contains values of all rows related to a particular field). Relationships between tables specify the association between parent and child tables, which can be one-to-one, one-to-many or many-to-many. Features such as integrity constraints and normalisation (reduction of data redundancy) improved the relational model by simplifying data management and data retrieval. As many databases fail to adequately support the normalisation process, which is an essential part of database design, the relational database remains the best option for most cases in business [137].

There are two types of relational databases which can be used based on the data usage requirements (transactional or decision making): the online transactional processing (OLTP) database and the online analytical processing (OLAP) database. OLTP databases are the most commercially available databases used to create transaction-oriented applications to manage operational data which can be used by thousands of users, and normalised tables serve simple to complex queries [178]. On the other hand, the main purpose of OLAP databases is to provide end-users with data in response to submitted queries. OLAP systems are designed to provide advanced data analysis to support decision making based on both operational and data warehouse data [136]. Data warehouses contain data collected from all parts of an organisation's OLTP systems to support management decision by providing a historic and summarised data view of significant information [137].

5.2.3.1 Database design lifecycle

The process of the database lifecycle starts by selecting a relevant approach to design the database from a set of business rules, processes and data that have already been defined [137], for example, *a data-driven approach*, *a process-driven approach* or a *parallel/blended approach*. The focus always remains on data, and business functions and rules throughout the development and implementation stages. Also, working within a structured framework for database development is essential considering some aspects such as goals and objectives, expectations, current and future requirements, implementation strategies, development tools and techniques [175]. Also, a design methodology must be selected, which ensures that aspects of database design are taken into consideration, resulting in a high-quality product.

Regardless of the design methodology selected, the design process involves essentially the same process of creating a logical model and converting it into a physical working model. The database lifecycle shows what steps are needed in a methodical approach to designing a database, and it includes the logical database design which is the process of modelling the information of the conceptual level using a specific data model [177]. The logical model is independent of the system environment of the physical design, which is based on the database management system appointed for the implementation [178].

Logical design: Throughout this stage, the focus is to obtain a logical structure design of the database (logical model) to support the functional requirements of existing and potential data and processes. The logical design methodology for the design process encompasses modelling requirements using the entity-relationship approach for data requirements specification and conceptual modelling [178]. During the logical design of the relational database, any entities of many-to-many associations will be further broken into two or more entities as all the relationships should be of a type of one-to-many or one-to-one. This process is called normalisation, which can be achieved by applying logical rules to the entities [136].

Physical design: This stage is about the database management system and hardware-dependent; it deals with the efficient data storage and retrieval process from the physical storage. This process affects the location of the data in the physical storage as well as the performance of the system [137]. The physical design deals with the conversion of the logical model elements' entities into database tables and the entities' attributes into columns as well as clearly and completely defining the constraints for the designed model. Also, through this process, the main objective is to provide an accurate representation of real-world data elements [178].

Database implementation: After the completion of the design stage, the formal schema can be implemented to create a database. The SQL, with both its parts, the data definition language and the data manipulation language, is used to build, query and update the database [178]. During this phase, aspects such as defining indexes, sizing tables, establishing constraints such as referential integrity and using available hardware components are undertaken. Then the initial schema should be tested after the generation of the data definition language and before database deployment [136].

5.2.3.2 Structured query language

Structured query language (SQL) is the standard language defined by the relational database community and standardised by The American National Standards Institute used for communication among relational databases. The SQL contains the data definition language to define the structure of the database; the data manipulation language to modify data within the database, and the data query language to easily retrieve information from the database [136]. Generally, the SQL can be used to query and update the database as well as to set up indexes and establish constraints, such as referential integrity [178]. A large number of relational database management systems follow the standard by using all keywords or extending the standard by adding new keywords [177].

5.2.3.3 Database management system

The database system encompasses software and hardware components. The database management system is a generalised software system for manipulating databases, and it provides all the required programming support, including scheduling of user programs, file management, database manipulation and error recovery [14]. During the physical design phase, the selection of the database management system should take into account its functions, features and facilities [175]. The most widely used database management systems are the relational database management systems as they provide a higher degree of data independence than the other database management systems (the hierarchical and the network). Relational database management systems support the three-schema architecture (external, conceptual and internal) of the modern relational database to separate programmers and end-users of the database from the physical storage of the data and present each user with the relevant subset of the data according to their needs [138].

5.2.4 NoSQL databases

NoSQL database refers to a non-relational or “not only SQL”, is a large distributed open-source system that provides a flexible and scalable environment for storing and analysing large volume of different data from multiple sources [179]. NoSQL database systems have recently attracted the attention of industry and researchers because of the demand for high-performance access to large amounts of data without the need for a large effort to expand and adjust. Since the NoSQL database technology relies on horizontal scalability, allowing increased performance and capacity by increasing the number of nodes, rather than increasing the computation power of a single node [180]. The rise of NoSQL technology was linked to large companies such as Google, Amazon, Facebook, etc. NoSQL database can be grouped into four categories for data modelling. The simplest version of NoSQL is *Key-Value stores*, where any data item can be a key to a stored digital object [181]. Key-Value Scales to unlimited data size, for example, is Amazon’s Dynamo that provides incremental scalability [182]. *Document stores* are the dominant version in the NoSQL databases [181]. Document stores are associated with the object-oriented programing throughout the entire process steps from clustering to accessing the data. Also, document stores have the same behaviour as key-value stores, as a value associated with a key is the document-content. They are useful for data with high complexity such as medical records; examples of this type are MongoDB and CouchDB [183]. *The column-oriented store* is a database which organized into related column group, this type inspired by Google’s BigTable, which is distributed, strong, a multidimensional sorted map [182]. *Big Table* was developed by Google based on the GFS to manage highly scalable structured data [184]. Big Table is a sequence of nested key-value pairs where keys and values can be composed as Apache HBase, and Cassandra, an open-source database management system [185]. In the context of NoSQL databases, programming languages have been introduced like MapReduce to minimize the complex tasks for data processing and reduce the performance gap among relational databases. As a result, programming models become a foundation for data-processing paradigm for highly scalable, fault-tolerant, large scale distributed applications [186].

5.3 Clinical data management

Electronic medical records contain all the information gathered on the health problems of a specific patient. However, clinical research databases are created for research purposes to accommodate selected information from all the patients’ clinical data sources, such as electronic medical records, sensors, implanted devices, in-home care devices and mobile devices, targeting specific medical problems or the technology under investigation. Consequently, clinical research databases tend to be specific to a disease, population, procedure, treatment or device. The automatic transfer of patient data between patient care databases and clinical research databases can help reduce data duplication and increase consistency [132]. The structural organisation of the clinical research database is designed to support data retrieval and answer research questions using software tools for custom queries, reporting and statistical analysis.

5.3.1 Data collection

Healthcare data can be obtained using primary data collection methods such as observations, surveys and interviews. In addition, secondary sources of data which are pre-organised healthcare data can be obtained from the electronic patient record, research articles, the internet. Both methods can be used for research and healthcare management. Although the primary data collection method provides unbiased, current and independent information, it is still a very expensive method and produces limited information. However, secondary data collection methods can provide unlimited data but with many concerns about data reliability and usefulness [12]. Hospital information systems are used to collect billing and clinical care data on a daily basis which may be suitable for administrative reporting but need pre-processing to extract useful knowledge for research and improve the quality of care [13].

The data collection process through healthcare systems may result in issues, such as incomplete, inconsistent and noisy data. Sometimes, physicians collect their findings using free-text notes, or by dictation, and these reports need to be transcribed into the computer. Data collected through this method should undergo the categorisation process into a group of functions such as diagnosis, treatment and plans [14]. Otherwise, the collected data can be difficult to manage for research purposes unless it goes through the preparation process to make it useful for research. The pre-processing data phase is the best solution to improve the quality of data which affects the analysis outcomes. Data pre-processing focuses on the preparation and transformation of the initial dataset. Therefore, this stage contains methods of data cleaning and noise handling, data integration and data transformation using a standard format, and finally, data reduction in summary reports.

5.3.2 Data pre-processing

Data collection and pre-processing are the most significant and fundamental stages by which to acquire correct and appropriate data for further analysis tasks. Data preparation is essential in discovering the required knowledge, especially from a field which generates high-volume data such as healthcare. Clinical care data which can be generated at various points of care might be structured, unstructured or semi-structured and comprise a wide variety of data types and formats, such as text, numbers, images and video. Therefore, knowledge cannot be acquired, comprehended and automatically extracted without the application of pre-processing techniques. The secondary use of extracted data from a health information system requires the use of pre-processing measures to eliminate data quality issues which may result from missing data or incomplete medical records [13]. An efficient and robust pre-processing algorithm needs to be implemented prior to data transformation and loading into the database [14]. For example, data cleaning techniques use methods to impute or fill incomplete data or treat noise either by polishing and correcting or filtering and removing the noisy instances [187].

5.3.3 Data storage

A data storage system contains two layers: a hardware infrastructure in the lower layer, and storage methods or mechanisms on the top layer. The hardware infrastructure is a combination of both hardware equipment such as servers, routers, network links and software components such as operating systems [188]. Data storage methods are on the top of the physical layer, equipped with application programming interfaces which enable rapid performance for queries over relational data in addition to another programming model for data analysis and better interaction with the underlying physical storage [184]. Current storage mechanisms can be classified into three bottom-up levels: file systems, databases and programming models [184].

5.3.3.1 File systems

File systems are the base for the applications at upper levels. The Google file system is an example of a highly scalable and consistent distributed file system for large-scale data-intensive applications [179]. However, it has some limitations, such as poor performance for small files and a single point of failure [184].

5.3.3.2 Database systems

Database systems have been developed over the past decades to manage various types and scales of datasets. Database technologies, such as data warehousing, have been used for big data storage for some time and have contributed to the development of several storage techniques [189]. In addition to relational databases, these include object databases, XML databases and multidimensional databases, which provide greater support for traditional datasets but are unable to meet the challenges brought by big data. Alternatively, NoSQL databases achieve greater performance than traditional relational database management systems [179]. The simplest version of NoSQL is key-value stores, where any data item can be a key to stored digital objects [181]. The key-value scales to unlimited data size, for example, Amazon's Dynamo, which provides incremental scalability [182]. Document stores are the dominant version in NoSQL databases [181]. Document stores are associated with object-oriented programing throughout the entire process from clustering to accessing the data. Also, document stores have the same behaviour as key-value stores, as a value associated with a key is the document content. They are useful for data with high complexity, such as medical records. Examples of this type are MongoDB and CouchDB [183]. Column-oriented stores are databases which are organised into related column groups and are inspired by Google's BigTable, which is distributed, strong and a multidimensional sorted map [182]. BigTable was developed by Google, based on the Google file system to manage highly scalable structured data [184]. BigTable is a sequence of nested key-value pairs where keys and values can be composed as Apache HBase, and Cassandra, an open-source database management system [185].

5.3.3.3 Database programming model

Database programming models have been developed to achieve effective distribution at scale for data-intensive applications. In the context of NoSQL databases, programming languages such as MapReduce have been introduced to minimise the complex tasks for data processing and reduce the performance gap among relational databases. As a result, programming models have become a foundation for the data-processing paradigm for highly scalable, fault-tolerant, large-scale distributed applications [186].

MapReduce is a powerful programming model for large scale applications which uses a simple technique that emerged from those used in the area of distributed databases [188]. MapReduce is a parallel programming framework developed by Google based on the Google file system for global analysis in big data [181]. The fundamental role of MapReduce is based on the divide-and-conquer method. In the "map" step, the programming task is divided into sub-tasks using the mapper function which takes the input as a key-value pair and distributes the smaller sub-tasks to be solved in a parallel and separate way. Then, in the "reduce" step, solutions from different distributed nodes for the sub-task are combined to provide a solution to the original task [190]. The MapReduce program can be written in a complicated low-level language such as Java which makes writing custom jobs difficult and time-consuming and requires a highly skilled programmer. Therefore, some advanced high-level query languages have been developed within the MapReduce framework, for example, Hive, Pig and Jaql [182].

Dryad is a programming model that implements parallel and distributed programs that are scalable and user-friendly [191]. Dryad's operational structure is a directed acyclic graph where a centralised job manager assigns computations to several processors, monitors the execution and is responsible for decision making [184]. Dryad is an independent system with complete functions that support job creation, monitoring, management and visualisation and also resource management, fault tolerance and re-execution [190].

5.3.4 Data analysis

Data analysis is the final stage of the data management process related to clarifying the meaning and understanding of the data collected and is organised for research purposes. Data analysis methods and techniques are applied for the interpretation of the results, writing reports and evaluation [192]. Descriptive statistics have been used merely to describe what has happened, such as in the most popular statistical package, SPSS. Also, past information predictive and prescriptive analytics are used to predict the future outcome and to direct future activities to achieve the best results, respectively [189]. The analysis of structured data reached an advanced state which now relies on a mature technology such as relational database management systems, data warehouses, or online analytical processing. The analysis is mostly based on a data mining and statistical approach in addition to statistical machine learning which has been applied to detect anomalies in the data using mathematical models and powerful algorithms [184]. On the other hand, unstructured data analysis, such as text mining, is a process of extracting useful information from unstructured text. Some text mining systems use a rule-based approach to identify patterns; however, others use machine learning techniques like natural language processing and other algorithms to discover patterns automatically from datasets [193].

Data analytics plays a significant role in highlighting the most useful data for diagnosis, treatment and discovery to improve the quality of healthcare by identifying data patterns and the relationships among them to develop more insight using algorithms and analytics tools [194]. Big data analytics has had a pervasive impact on healthcare which is clearly visible in different areas, such as improving the efficiency and quality of care while lowering the cost, the early detection and prevention of disease, and fraud detection by automating the verification of claims to prevent fraudulent claims [195]. Although healthcare systems have all the requirements for the effective application of big data analytics, such as data-intensive and critical decision support, challenges such as interoperability issues and privacy and security concerns remain [196]. As stated in the 2011 McKinsey Global Institute Report, big data analytics can effectively contribute to different areas such as clinical operations, research and development, and public health to provide a better outcome and reduce waste and inefficiency [49] [185].

First, within clinical operations, outcomes-based research such as comparative effectiveness research determines the most relevant and cost-effective treatment for a patient depending on the analysis results from the comprehensive patient and outcome data. Also, the use of clinical decision support systems helps reduce the number of clinical care mistakes, reduce treatment error and adverse reactions, and enhance the efficiency and quality of operations [185]. The implementation of advanced analytical methods, such as segmentation and predictive modelling on patient profiles, identifies patients at risk who may benefit from proactive care or lifestyle changes [49] [185]. Furthermore, the use of evidence-based medicine for the detection and prediction of at-risk patients based on big data gathered from various healthcare sources provides sufficient evidence to identify and deliver effective clinical care [197] [194].

Second, in R&D, predictive modelling has had an incredible impact on disease diagnosis and treatment [198]; it not only leads to the prediction of clinical outcomes and new drugs but also includes evaluation factors such as safety, efficacy, possible side effects and the final trial outcomes. The clinical phase of the R&D process can benefit from the application of statistical tools during patient recruitment to improve the design of clinical trials as well as to analyse clinical trial data and patient records. This will identify further signs and discover adverse effects as well as enable the detection of rare safety signs that appear in a typical trial and reduce drug withdrawal from the market [185].

Third**,** in public health surveillance and response, analysing a nationwide patient and treatment database for the rapid detection of infectious diseases and outbreak provides a quick surveillance response and reduces infections. These analyses can also be used for the rapid development of more accurate targeted vaccines [49].

The healthcare industry can exploit diverse data analytics technologies to process and analyse medical data to improve healthcare services. The two widely used techniques in using such data are information retrieval and data mining [50].

Information retrieval is the most commonly used technique that deals with the process of "acquisition, organization, and searching of knowledge-based information" (p.467, [198]). Information retrieval can be used to obtain information by searching for a specific user's query within a large document collection where the retrieved subset of information is in the same format as the original with no added values [50]. Traditionally, information retrieval focuses on the retrieval of text from medical data; however, now it covers a wide range of digital media, including the retrieval of medical images [198]. Medical text retrieval can be considered to be a domain-specific text search with the significant challenge of dealing with the inherent complexity and ambiguity of medical terminologies that require standardisation. Therefore, semantic-based text search approaches are used to tackle the ambiguity issue in medical text. However, for medical image retrieval, either text-based or content-based approaches can be used. Text-based retrieval depends on the annotated text associated with images, while in content-based medical image retrieval, the process depends on the description of the visual features of the image such as colour which can be automatically generated while indexing the medical image [50].

Data mining is the process of extracting patterns from massive datasets by combining methods from statistics, machine learning and artificial intelligence with database management systems [185]. Healthcare data mining concentrates on comprehensive questions and outcomes, for example, symptoms and all the related data and clinical outcomes in combination lead to particular diagnoses and treatments [196]. The application of data mining in healthcare can be classified into supervised (predictive) and unsupervised (descriptive) approaches. Supervised learning methods are used to build clinical prediction models based on predicting a function or associations from a set of training data [185]. These methods have been successfully used in clinical prediction using statistical methods (i.e. linear regression, logistic regression and Bayesian models), sophisticated methods in machine learning and data mining (i.e. decision trees and artificial neural networks) and survival models that try to predict the time of the occurrence of a specific event. Generally, supervised learning methods can be classified into two broad categories: classification and regression, where both focus on discovering the underlying relationship between covariate variables and a dependent outcome variable [198].

Unsupervised learning methods which involve data clustering find hidden structures in unlabelled data [185]. These methods depend on grouping data into clusters according to the objects' (patients' or health records') similarity measurements. Examples of unsupervised or descriptive data mining approaches include clustering, association rule mining and sequence discovery [50].

5.4 Clinical research databases

Clinical research databases can be primary databases where data are collected specifically for research, such as clinical trial studies. However, generally, they are secondary databases that contain a specific group of data extracted from primary databases such as electronic medical records with a common problem. Clinical research databases can be categorised according to their analytical purpose into (i) descriptive analyses to extract summaries of the essential features of a database, such as grouping patients with similar conditions, and identifying the critical characteristics of each condition; and (ii) predictive analyses to derive classification rules, such as developing diagnostic standards which predict the course of a disease. Clinical research databases require de-identifying all patient data before including and using linking variables to link patients who may be related to more than one source, but the patient's privacy and confidentiality always need to be maintained. The U.S. Department of Health and Human Services enacted the Health Insurance Portability and Accountability Act privacy rule that allows the use of healthcare data after removing an individual identifier [199].

5.5 Structural designs for clinical databases

Digital technology has grown rapidly in the healthcare sector, leading to a significant shift from paper to electronic records, thereby increasing the volume of healthcare data which, as a result, requires databases to manage, manipulate and store. Databases are fundamental to the effective use of data to serve an organisation's multiple purposes [14]. The conceptual representation of an individual patient and modelling the schema for patient care during the healthcare process can be done using any structural design such as hierarchical, relational or object-oriented, or a hybrid structural design. The design should adhere to the basic functional requirement to serve the primary goal of the medical database: (i) provide easy access to all relevant data for each patient served; and (ii) provide a resource for the scheduled retrieval of all relevant data from the records of all patients for any primary or secondary purpose [199].

The relational database is the most common database used in the healthcare system, which can be in the form of administrative and billing data or recording patient care, surveillance health status and treatment advice. In addition, it can be used for research purposes to assist researchers with studies such as drug effectiveness and diseases [200].

There are various data storage models under the relational database concept such as the entity attribute value model which is the most widely adopted storage model in clinical systems. It has a three-column fixed schema of an entity, attribute and value which are used to store the primary key, the attribute name and the data value respectively. The entity attribute value model improves flexibility by allowing attributes to be added by simply specifying their names in the attribute column. However, the main model drawback is the restriction on a single value column which hinders the ability to use multiple data types [134]. The type most commonly used in healthcare is the online transactional processing database, which can contain applications such as electronic health records, administration, billing and payment processing, financial systems, HR and research. It provides real-time transactional processing (search, store, update, delete) with the fast response time. In addition, the online analytical processing database, which is a data warehouse, can be built on the top of existing multiple online transactional processing databases to combine data with analytic purposes [201].

Design challenges

There are many challenges that a database designer has to overcome, such as the structure and relationships of unique and complex healthcare information. For example, demographic information can relate to multiple diagnoses which, in turn, are linked to other elements, such as the procedures performed by many doctors who can prescribe many medications [200]. Although there is a constant change in information requirements due to advances in medical fields, historical data remains valuable and does not diminish like the majority of data in traditional business information systems. During the design stage of the information system lifecycle, special attention is required for the medical field requirements, such as the common vocabulary of generally accepted terms for medical concepts and for administrative data used for patients. The issue of data sharing in integrated applications is very important and is affected by many factors, including lack of adopting technological solutions for integrated and interoperable sharing of data [202].

Modern healthcare depends on collaboration and communication. With the growing application of health information interchange systems, there is a need for interoperability to provide information when and where necessary, facilitate decision making, reduce waste by eliminating redundant work and improve safety with fewer errors. Interoperability can be seen as four layers of “technology, data, human and institutional” with corresponding types of interoperability “technical, semantic, operational and clinical” [203].

- *Technical interoperability* is the technology layer, where information can be exchanged by health information technology systems without any ability to interpret the data. This foundational layer is domain-independent, as reliable communication can be achieved over a noisy channel.

- *Semantic interoperability* is the data layer, where health information technology systems exchange, interpret and use data without ambiguity. But this layer is domain and context-specific, which requires the use of standardised unambiguous codes.

- *Process interoperability* is the human layer, where process interoperability is achieved when people share a common understanding of their process artifacts across the network.
- *Clinical interoperability* is a subset of process interoperability, which is the ability of two or more physicians in different care teams to transfer patients and provide smooth patient care.

As health data standards are a necessary component of interoperability in healthcare, poor implementation of interoperability results in failed large investments in digital health. The application of standards in healthcare will enhance the interoperability of healthcare systems to deliver timely services and provide better healthcare to patients [204].

However, integration is generally viewed as going beyond mere interoperability to encompass a certain degree of functional dependence [202]. Data integration is a prerequisite for obtaining a unified view of clinical and genetic data from multiple operational data stores for a different healthcare organization. Integration of data stored in heterogeneous DBMS can be achieved by aggregating across data sources using an intermediate solution to retrieve data or combine copies of data in a centralized system from multiple sources to provide a unified view using various techniques for data extraction. The process of data integration involves extracting, transforming, and cleansing the data before aggregation. Several approaches have been employed to integrate data from different sources using warehousing approaches also known as physical integration (combining copy of the data in a new repository for further analysis) or using mediation-based approaches known as logical integration(applying conceptual schemas to bridge representational heterogeneity of the databases providing queries with the ability to collect and integrate data from distributed sources) [18].

5.6 Supportive technologies for clinical research database integration

Health-related data capture, storage, analysis and retrieval is rapidly transforming from the paper-based system to the digital system. However, the sheer size, as well as the complexity of this data, make it difficult to process and analyze the data through traditional methods and techniques. Hence, technologies like cloud computing and NoSQL are now gradually being used to integrate and process massive data effectively and securely in healthcare [109].

Cloud computing

Cloud computing is one of the powerful technological advances that has emerged in modern ICT. In cloud computing data, scalable computing resources and other services are provided over the internet at a lower cost [185]. Cloud computing provides services such as virtual resources, parallel processing, data integration and scalable data storage.

NoSQL databases

Although relational database architecture provides numerous advantages such as high consistency and availability, its performance decreases as the data grow and it faces scalability constraints as it is impossible to scale horizontally and its vertical growth is limited. NoSQL databases provide solutions for data aggregation and handling unstructured data, and its schema structure is flexible. In addition, it provides scalability for a quickly growing data repository, where horizontal scalability is one of the NoSQL databases features [160].

5.7 Summary

The design of a database depends on the purpose of the database and its use, and this helps determine how the design begins, the type of model to be chosen, and how the database will be implemented and managed. The concept of database design has been explained in detail in this chapter to emphasise the importance of the design process in producing a transparent model that ensures data integrity and adheres to the organisation requirements as well as delivers simplified and easy to perform databases. The complexity, rapid development and expansion of the clinical information field makes it difficult to develop and maintain clinical databases [205]. The design and implementation of a clinical research database need more attention paid to the requirements to understand the conversion from an abstract model to a functioning database. To use databases in research, the relational database is the best option to allow more search options that support complex queries for research questions and allow easy reporting options for novice users. However, for larger healthcare organisations, clinical data warehouses provide a comprehensive view of integrated data to support large studies as well as clinical decisions. Immersive technologies such as cloud computing and NoSQL databases can be used to integrate data from disparate sources with different formats and structures into a unified repository.

This chapter investigated some theoretical considerations for designing and implementing a database for clinical research to achieve the required specification, integration capabilities and user satisfaction for a data management system. Next, in Chapter 6, the architectural framework of the solution is displayed based on the solution framework and the theoretical consideration in the design of such a system.

Chapter 6: Architectural Framework for the Proposed Solution

6.0 Chapter overview

This chapter demonstrates the development of the architectural framework of the proposed solution. Section 6.1 presents the purpose and objectives of this chapter and provides connections to previous chapters. Section 6.2 outlines the design objectives, while Section 6.3 notes the characteristics of the data management system G3DMS, and Section 6.4 presents the characteristics of the integration framework and the resulting GENE2D integrated data repository. Section 6.5 presents the solution architectural framework with a brief description of both the G3DMS and the GENE2D. Section 6.7 overviews the thesis solution methodology to achieve the intended goal. Finally, Section 6.8 summarises the chapter and introduces the next chapters.

6.1 Introduction

This chapter represents the last section of the problem analysis stage and the first section of the design stage in the abstract world (Figure 1.1, Figure 1.2), the preparation of the conceptual representation of the solution in a distinct architectural framework. The outcomes are from the problem analysis stage which includes the literature review in Chapter 2, the pilot study in Chapter 3, the problem definition and the solution framework in Chapter 4, and additional theoretical investigation in Chapter 5. This chapter focuses on the development of the architectural framework of the solution based on the outcomes derived from previous chapters.

Preliminary

The Saudi government supports many projects and programs that aim to collect genetic data to help research and develop preventive measures to limit the prevalence of these diseases in the region. Genetic clinics and research centres have helped in revealing genetic disorders and generating curated data which can be a valuable source for researchers if the data are collected properly and stored efficiently. Sharing such data could advance research and enhance healthcare quality. However, these centres lack adequate systems for managing and archiving patient data for research, making it difficult to integrate and build a national genetic diseases database. Internal problems within Saudi clinics and genetic research centres in terms of data collection and retention for use in research hinder data integration and sharing for research and the establishment of a unified database of genetic disorders.

This thesis contributes to identify, define and propose a solution to the problem through its methodology and research approaches provided in previous chapters. The problem statement clearly states the issue in these clinics and research centres in two main points: (i) lack of an efficient data management system for care and research; and (ii) lack of an integration framework between these centres. These prevent a unified view for genetic studies and the generation of a Saudi specific genetic disorders database. Accordingly, the thesis proposes a twofold solution:

- the design and implementation of a novel genetic disorders diagnosis data management system called G3DMS
- the design and implementation of a NoSQL-based integrated data repository for genetic disorders data called GENE2D.

6.2 Design objectives

The development of the information system, the database and its application interface would help to improve and solve problems related to data provision, integration and exchange for medical research besides the improvement of healthcare quality and outcomes. The thesis plans to provide a solution that addresses all the concerns in the current system issues and provides new characteristics that support health informatics principles and leverage the healthcare workflow, enhance decision making, and make data reusable for research.

- ***Data quality:*** The system interface and database design consider attributes such as data *completeness*, *accuracy* and *consistency*. The use of standardised forms for data entry with validation action will ensure and enforce certain aspects of data quality to obtain information reliability to serve the purpose of research.

- ***Data availability:*** The system ensures availability in terms of accessibility and durability, and the system grants users access to the database when and where needed and ensure data are valid for efficient information retrieval.

- ***Data integrity:*** The system ensures data are properly recorded as intended, and pre-defined rules are consistently applied to all data entering the system. For example, it implements checks data before performing any CRUD operations (Create, Read, Update and Delete).

- ***Ease of use:*** A researcher with average computer skill can use the system and thus can collect research data and achieve their study results without the assistance of programmers or the need to know the complex query language. Easy to use dropdown menus that reflect the actual information from the database encourages researchers to research their locally collected data, which gradually can source more substantial datasets.

- ***Provide a working environment regardless of any difficulties:*** The system permits offline patient data collection using an Excel template, later to be uploaded as bulk data all at once to the database. The system is accessible from any device via a web interface, including smart devices.

- ***Support legacy data migration:*** The system allows legacy data migration using customised mapping method to upload an Excel template to the new database.

- ***Facilitate the process of clinical decision making:*** The web-based system allows real-time access to patients' data with all related supporting photos, and documents such as reports and publications which helps physicians to make an informed decision.

- ***Maintain data privacy and security:*** According to management policies for data privacy and security, the system can be implemented on the organisation's local server or on a hired private web server. User access authorisation can differentiate the role of regular users from the administrator in order to preserve the privacy and security of information.

- ***Increase research productivity:*** The system's ability to combine multiple sources of patient data into a single cloud-based repository provides significant opportunities for research to be performed on large datasets that increases the efficiency and productivity of research in the region.

6.3 Characteristics of the novel G3DMS

The design objectives identify the features that should be incorporated into the design and implementation of the data management system for supporting genetic disorders diagnosis workflow and use in research, including:

- electronic data capture forms for new data entry

- interface for uploading data from old systems format (Excel files)

- interface for data modifications (update, delete)
- interface for reporting data (display individual patient and cohort)
- interface for customising queries for research
- interface for extracting unidentified data for research purposes in other formats (Excel, JSON)
- provide secure multi-user access (authorised users).

6.4 Characteristics of GENE2D

Besides the design objectives, there are design considerations such as the solution should suit low-resource settings and limited information and communication infrastructure. Therefore, for the integration framework to be valid in these conditions, the GENE2D should be delineated by the following features:

- be cost-effective and easy to implement at the clinic site
- provide simple and convenient methods for moving data from G3DMS to the GENE2D
- provide a cost-effective exchange, transform and load process
- provide a de-identification method at the source G3DMS before moving data for integration
- provide a query interface for customising simple and complex query operations
- provide secure multi-user access (authorised users) to the GENE2D.

6.5 Proposed solution framework

Figure 6.1 shows the architectural framework of the final product of the proposed solution. The components of the solution include a standalone new genetic disorders diagnosis data management system (G3DMS) to be implemented in any genetic clinic or research centre; and an integration framework in the G3DMS and a NoSQL-based integrated data repository of genetic disorders data (GENE2D). Figure 6.1 shows the integration framework of multiple sources with G3DMS (clinic and research centres) which regularly send selected data from their databases to be aggregated into the GENE2D where all combined data for research can be accessed by authorised users through a web interface. Figure 6.2 displays the system architecture of the G3DMS which includes a web interface for managing the process of the genetic diagnosis including standard forms for data collection, manipulation, retrieval, reporting and analysis as well as uploading legacy data and downloading research data in various formats.

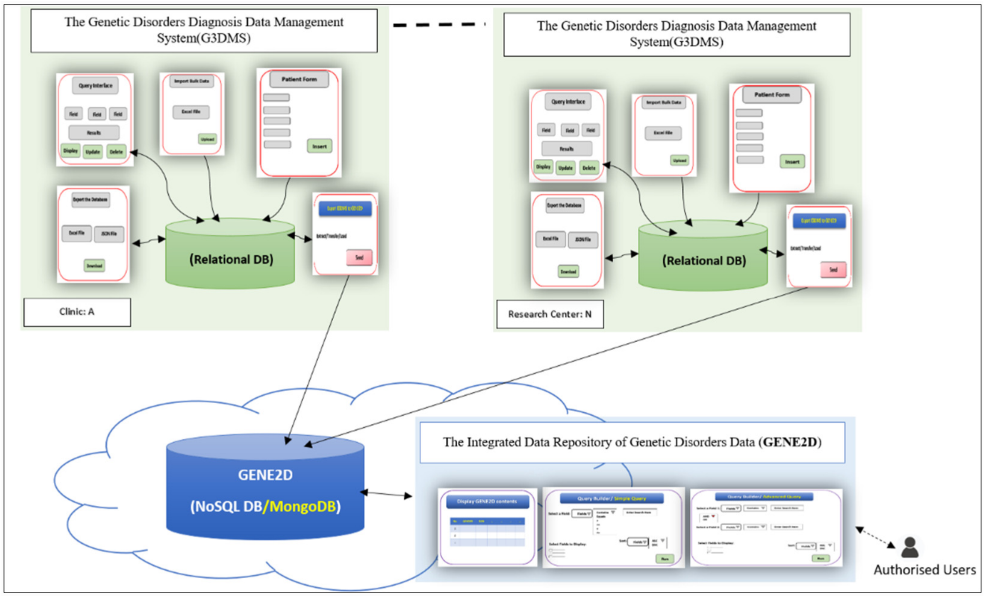


Figure 6.1: Proposed solution framework

6.5.1 Genetic disorders diagnosis data management system (G3DMS)

The source for the integration framework in Figure 6.1 is the G3DMS. The effective operation of the integrated framework requires the data sources to be valid and to have the ability to contribute effectively to the operation as well as provide quality data for research. Figure 6.2 shows the G3DMS architecture as a standalone system and a component for participating in the integration framework as a source of data. First, the standalone G3DMS contains a web interface with electronic data capture forms for standardised entry, import data, export data and the query interface. The additional feature which makes the G3DMS a significant source for the integration is the ability to export directly to GENE2D. Pressing the send button will trigger multiple customised functions for data extraction transformation, and for loading to GENE2D directly without any user intervention.

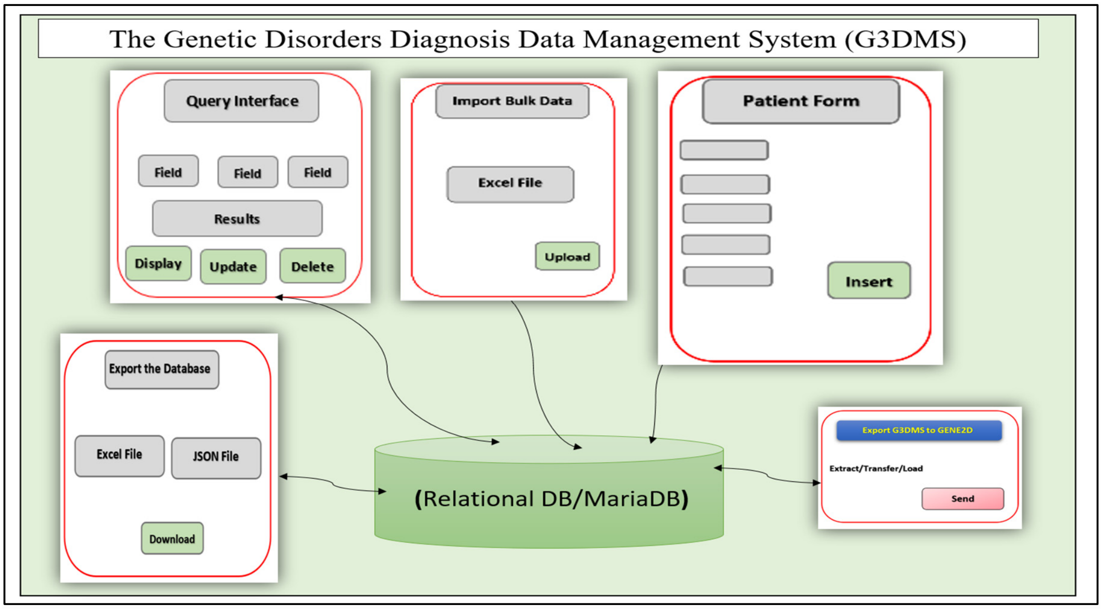


Figure 6.2: G3DMS architecture as an integration source

6.5.2 G3DMS architecture

G3DMS design objectives represented in the requirements are for both daily care for patients with genetic disorders and research related to this matter. Since the two tasks are interrelated and dependent on each other, it is essential to take into account their requirements during data collection, processing, archiving, speed and accuracy of information retrieval and the ability to link them easily to extract high-quality research outcomes. The major components of G3DMS are electronic data capture forms for data entry; a customised query builder to display and modify patient data as well as form research queries; a module that allows historical data to be uploaded in the form of bulk data using a template; export data options to Excel and JavaScript Object Notation (JSON) format; and authorisation access for healthcare researchers and clinicians. Additional features for integrating the data from the G3DMS into GENE2D include mapping methods for data extraction, transformation and loading.

6.5.3 Integrated data repository of genetic disorders data (GENE2D)

The NoSQL integration framework combines data on genetic diseases from multiple sources into the GENE2D system and provides an easy-to-use query interface for researchers to conduct their studies on large datasets. The major components involved in the GENE2D architecture consist of the data sources, the integrated data repository as a central database, and the application interface. The integrated data repository uses a NoSQL document store via MongoDB (an open-source document-oriented database program) as a backend database. The application interface called Query Builder provides multiple services for data retrieval from the database using a custom query to answer simple or complex research questions. The GENE2D system demonstrates its potential to help grow and develop a national genetic disorders database in Saudi Arabia.

6.7 Proposed solution methodology

The architectural framework for the solution in Figure 6.1 is executed as follows. First, design and implement the G3DMS in a full development process using a case study and the Barker method for guiding the systems development lifecycle, followed by a qualitative evaluation based on health informatics principles and user experience and satisfaction. Next, incorporate an integration framework into the G3DMS by adding features and methods for exchange, transformation and loading. Then use a physical integration approach to build an integrated data repository based on a NoSQL document-oriented database for flexible schema and cost-effective integration to be the storage mechanism for the GENE2D.

6.8 Summary

This chapter re-stated the problem which exists in Saudi Arabia, then outlined the design objective, which had been set to meet the criteria. The system architecture was depicted to show the whole framework of the proposed system (end-system). A detailed system description of the design and implementation of the G3DMS and the GENE2D systems was then provided to show the actual system components and parts. The thesis solution methodology was delineated to present the two execution steps with their fundamental design steps. The following thesis chapters convert the architectural framework to a working system. The design and implementation of G3DMS are presented in Chapter 7 and Chapter 8, and the integration framework and the design of GENE2D are demonstrated in Chapter 9.

Chapter 7: Genetic Disorders Diagnosis Data Management System (G3DMS): Strategy, Analysis and Design

7.0 Chapter overview

This chapter demonstrates the design process of a novel genetic disorders diagnosis data management system (G3DMS) based on the proposed solution framework in Chapter 4 and the architectural framework in Chapter 6. Section 7.1 introduces the purpose and objectives of this chapter and provides links to previous chapters. Section 7.2 presents the G3DMS description, while Section 7.3 describes the methodology adopted for the design of the G3DMS, using a case study following the system development lifecycle using the first three phases of the Barker method: strategy, analysis and design. Lastly, Section 7.4 summarises the chapter and introduces the following chapter.

7.1 Introduction

This chapter presents the last step in the design stage and the first step in the implementation stage of the solution in the abstract world (Figure 1.1, Figure 1.2). The purpose of this chapter is to deliver the actual conceptual and logical representation of the solution and answer the research question, see Figure 1.1, Q5.A: *How can a data management system be designed for a genetic clinic and research centre considering health informatics framework?*. The proposed architectural framework for the solution in Figure 6.1 is transformed into a working system in this chapter and the following two chapters. This chapter presents the design of the standalone G3DMS and produces the conceptual and logical models based on the strategy and analysis phases of the Barker method adopted for the system development lifecycle and the database lifecycle. The product of this chapter is used as a blueprint for the G3DMS construction and implementation in Chapter 8.

7.2 G3DMS architecture

The purpose of the G3DMS is to manage patient data collected during the diagnosis process to be used for decision making and research studies. Therefore, the structural components of the G3DMS are defined to serve the objectives of data collection, data storage and data reporting, where data are collected and processed for both an individual patient or a group of patients, taking into account a standard format for entering new data and transferring data from old systems. Figure 7.1 shows the components of the G3DMS: a relational database MariaDB as a backend for data storage; and the application interface, which includes a web-based interface with authorised access to the following features:

- electronic data capture to standardise data entry of patient information
- import bulk data in Excel format and map the content of the file to the database using a customised mapping method
- query interface to display, update and delete patients' records, and build customised queries
- export the content of the database to an Excel file and JSON file format.

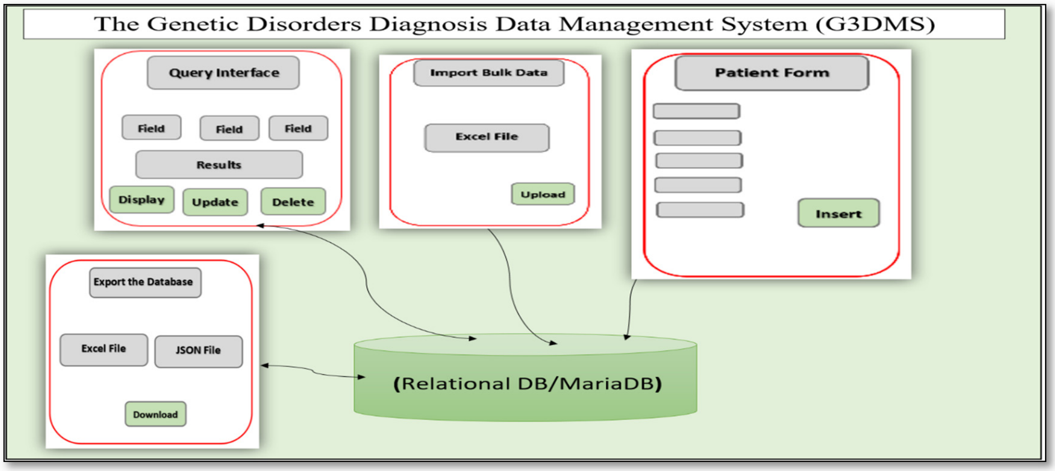


Figure 7.1: Genetic disorders diagnosis data management system (G3DMS) architecture

7.3 Design methodology

The system architecture in Figure 7.1 is implemented using two steps: first, a case study from Saudi Arabia is used to understand and analyse the current system limitations and propose a new system design model; and second, the design and implementation of the proposed system are examined using the Barker method for the full system development process.

7.3.1 Case study

7.3.1.1 Aim of the case study

This case study has two objectives: analyse the health information system of a genetic research centre in Saudi Arabia, the Princess Al-Jawhara Centre of Excellence in Research of Hereditary Disorders (PACER-HD) at King Abdulaziz University Hospital, Jeddah, and use the findings to propose a system design that could enable both clinicians and healthcare researchers to collect, manage, store and use genetic data for decision making and research.

7.3.1.2 Description

The PACER-HD provides several free medical services related to genetic disorders, including the Genetic Disorders Clinic which receives referral cases from different departments of King Abdulaziz hospital and from neighbouring hospitals in the region and neighbouring areas, where genetic cases are diagnosed at all ages; the Down Syndrome Clinic which is the only clinic in Saudi Arabia that provides a full examination, genetic counselling and follow-up in only one visit, and links patients to rehabilitation centres; and the Genetic Counselling Clinic which provides detailed genetic counselling services to families on consanguineous marriages, families planning preimplantation genetic diagnosis and families requiring genetic analysis of samples. In addition to medical services, the PACER-HD also established a laboratory to generate stem cells to conduct research and answer questions in the field of biological studies. The PACER-HD also has a molecular biology laboratory which is run by research teams for various projects and adheres to international standards and uses the latest techniques such as sequencing and DNA/RNA extraction to obtain accurate results and a cytogenetics laboratory which performs various analyses and experiments such as chromosome analysis, microarray, cell culture and the FISH technique [206].

7.3.1.3 Methods of gathering requirements

A focus group was held to brainstorm with the director of the centre and the clinicians participating in the research who were able to define the process of diagnosing genetic disorders to gather the requirements of the processes performed in PACER-HD and determine the data required to support the process. This helped perceive the problem from the perspective of healthcare professionals who communicated their problems and their needs directly without any ambiguity.

7.3.1.4 Problems with the current system

Based on the brainstorming discussion with the Director of PACER-HD and the professional staff, responses are summarised. First, the present system lacks the basic capacity to handle a simple research question on genetics. Second, although the available system allows researchers to make a single selection for a specific condition, it does not provide even a simple query answer involving more than one field (vertically) as the patients' data are stored in a flat-file structure with no relation between the fields. Third, the data gathered from paper-based documents and patient information are manually entered into the system by researchers using traditional Excel spreadsheets.

7.3.1.5 Current process model

The basic process model at the PACER-HD is presented in Figure 7.2, in which the first step is to collect the demographic information, the clinical phenotypes and family history. The next step is to make a diagnosis to determine if the case is identified. If the diagnosis is confirmed on the spot, the patient undergoes a treatment plan directly. Otherwise, a diagnostic test is requested to confirm the decision. The patient must give their consent before the sample is taken for the diagnostic test. The results are delivered to the clinic, and if the diagnosis is confirmed, a treatment plan for the patient is prepared. All this data must be recorded for every patient in the system of the PACER-HD with any additional documents or literature useful for diagnosis or research.

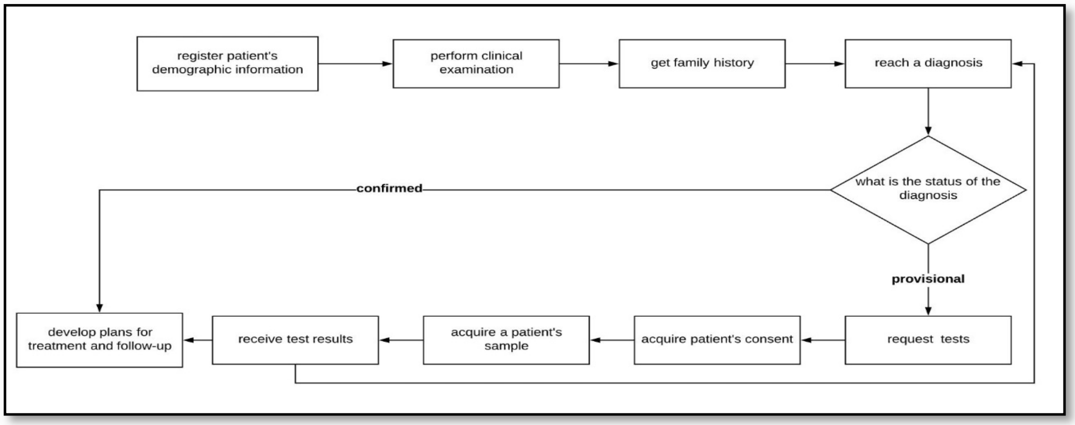

register patient's demographic information
perform clinical examination
get family history
reach a diagnosis
what is the status of the diagnosis
confirmed
provisional
request tests
acquire patient's consent
acquire a patient's sample
receive test results
develop plans for treatment and follow-up

Figure 7.2: Basic process model for a new patient's diagnosis

7.3.1.6 Data elements

The decision to include various data elements was made in discussion with the experts in this field, the PACER-HD director and the professional staff, to determine the essential fields required for the system and research studies. They summarised the required data elements they wished to include in the prospective system. Table 7.1 shows the data required by the centre, such as patients' clinical and demographic data, non-genetic investigation results, and the genetic testing results. The observation data is not applicable in this case as the local laboratory in the genetic clinic only collects patient samples for genetic testing to be shipped abroad to a diagnostic laboratory, and later receives the results and hand-delivers these to the patient's physician.

Table 7.1: Fields in the Excel file used for data collection of the clinical diagnosis process

Field Name	Description
MRN	Patient’s medical record number used to link patients to their hospital information
GN	The genetic number is given by the Centre for tracking purposes
NAME	Patient’s name
DOB	Patient’s date of birth
GENDER	Patient’s sex
ORIGIN	Patient’s nationality
DIAGNOSIS	Physician clinical decision of condition according to the patient’s symptoms and signs
IS IT DEFINITE?	Defines the status of the diagnosis if it is provisional or confirmed
INHERITANCE PATTERN	Shows if it is Autosomal Dominant (AD), Autosomal Recessive (AR), X-linked, etc.
CONSANGUINITY	Presents the existence of consanguinity for the patient’s parents (+/−)
MOTHER AGE	Age of the mother at the time of the patient’s birth
DNA-AVAILABLE	States the availability of the DNA test results (Yes/No)
s/b	Seen by or referred to the physician/clinician who is responsible for the patient’s treatments
NON-GENETIC INVESTIGATION	Clinical tests requested to confirm the diagnosis
GENETIC INVESTIGATION (RESULT)	Genetic test results to confirm the diagnosis
PLAN	Treatment plan such as follow-ups/rescheduling/referral
PHOTO	Patient’s photos (images)
CONSENT	Signed consent in a PDF format from patient
IMPORTANT NOTES	Additional publications/papers/survey in a PDF format related to the patient’s condition
CONTACT	Patient’s contact mobile number

The genetic test types that can be used for the diagnosis are provided by the laboratory staff at the PACER-HD, as shown in Table 7.2. Usually, patients’ data can be provided in the form of Excel files, and the test results can be submitted in portable document format (PDF) or image format (JPG).

Table 7.2: List of the types of genetic testing requested in the genetic clinic

Test Type	Test Name
Chromosomal testing	- Chromosomal analysis (karyotyping) - Chromosomal breakage - Fragile X chromosomes
Biochemical tests	- Metabolic screening test (urine or blood) - Enzyme assay - Serum amino acids
Molecular testing	- FISH analysis (for microdeletion or any specific loci) - A DNA Microarray - Methylation status analysis - Sequencing for a specific gene - Whole exome sequencing (WES) - Whole-genome sequencing
Testing for blood disorders	Hb electrophoresis
Preimplantation testing	PGD for single gene disorder or for specific chromosomal abnormality

7.3.2 Barker design methodology

The thesis adopts the Barker design methodology in the G3DMS for the full cycle of the system development process and the design of the database. Both the process models for the application interface design and the data models for the database design are considered during the design lifecycle. The Barker method extends the traditional method (requirement analysis phase, data modelling phase and normalisation phase) with more steps to better organise the design effort. The Barker method has seven stages, as illustrated in Figure 7.3: Strategy, Analysis, Design, Build, Documentation, Transition and Production [137]. At the strategy phase, the focus is on the data and application requirements depending on the process model and the data elements of the case study of the PACER-HD where the proposed system will be implemented. During the analysis phase, the requirements for the clinical genetic research database are determined by investigating the research requirements in terms of data, queries and processes. The analysis results are displayed using the entity-relationship diagram and show the detailed entity attribute relationships in the global schema, as well as the data flow diagrams in multiple levels for system processes. In the design phase, the logical schema, which was converted to tables and references, as well as the physical process flows which were converted to wireframe diagrams, are prepared based on the outcomes from the strategy and the analysis phases. In the build phase, the logical schema developed in the design phase is used to construct tables and references, and the application interface prototype is coded into actual web pages. Documentation of all stages of database design and application interface can be extracted from the diagrams and structural drawings in addition to the preparation of a user guide for system use. The transition phase encompasses the validation testing of both the database and the complete system in the development environment before deployment. Finally, the production phase includes the implementation of the system online to be accessed by PACER-HD end-users after moving the organisation's legacy data.

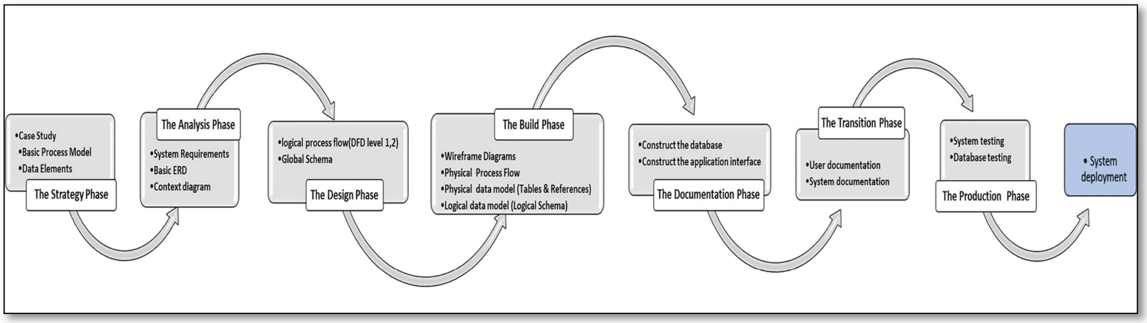


Figure 7.3: Design lifecycle of the G3DMS using the Barker method

7.3.2.1 Strategy phase

The design and development of a data management system for an organisation require consideration of all the information involved in the business processes. Therefore, the requirements definition is an integral step in the data modelling process to specify the data content of the database [175]. Hence, a customised strategy is used to design a data management system for the diagnosis of genetic disorders using the major process that results in generating patient data that can be useful for research if captured and stored appropriately. Therefore, the strategy is to cover the following points using appropriate sources from the PACER-HD case study:

- highlight the current services provided by the genetic clinic or research centre
- identify the processes performed in the centre and understand the tasks and the role of the researcher and the effectiveness of the current system in meeting the research process requirements
- identify the data elements required to perform the processes
- determine the research process requirements and the researchers' expectations of the prospective system for data collection, management and analysis.

Context diagram (data flow diagram level zero)

The basic process model for the primary process performed at the PACER-HD is illustrated in Figure 7.2. This shows the main entity is the patient, and the primary user of the system who performs all data entry into the system is the physician. Therefore, the next step in the development of the process models is to draw the data flow of the diagnosis process identified in the case study. Figure 7.4 shows the context diagram or level zero data flow diagram, which graphically represents the whole system as a single process, emphasising the interaction of the external entity or the physician with the system. All requests and orders from the physician to the system are presented as an input operation and the system response as an output operation. The data flow level zero diagram is foundational to the development of the logical flow of data through a system to perform any task required from the system, for example, add new patient, display patient information and delete patient records.

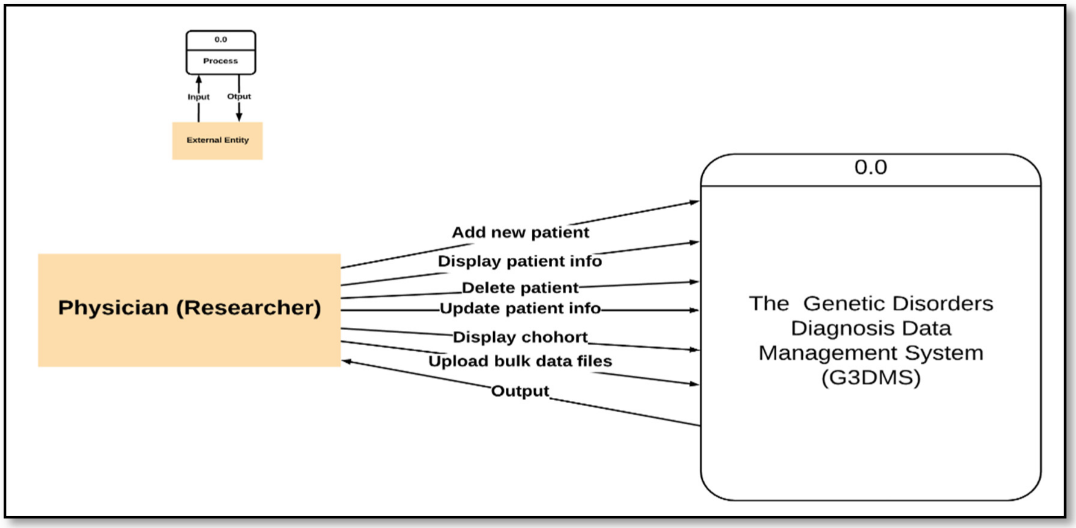


Figure 7.4: Context diagram – data flow diagram level zero

Basic entity–relationship diagram

The data elements (entities) are derived from the diagnosis process elements and data required by the PACER-HD in Table 7.1. Entities or main objects that play an essential role in the diagnosis process and their primary relationships are described in the basic entity-relationship diagram. Figure 7.5 shows the basic diagram, which only displays the connection between objects without emphasising the degree of the relationship. The patient is the central entity in the design, as all other entities must have a direct link with the patient. For example, the patient must have demographic information and family history; however, the patient may have several conditions, clinical phenotypes, tests, samples and treatment plans.

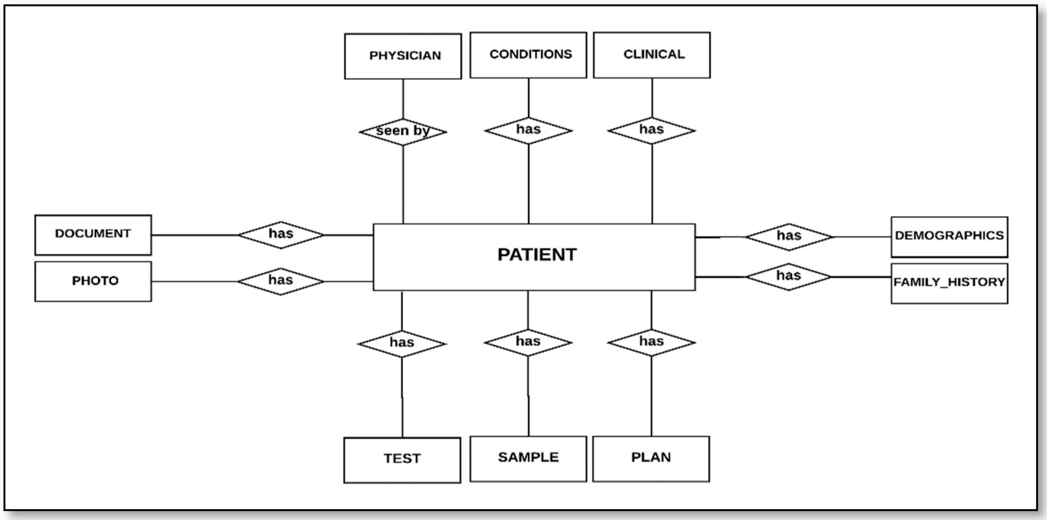


Figure 7.5: Basic entity–relationship diagram

System requirements

The focus of the strategy is to use the data from the case study and extract a plan for preparing a process model and data model in addition to the system requirements which are discussed according to the current issues with the PACER-HD system. Determining the expected system requirements in terms of the rules governing data entry and storage in the database will help define the requirements of the database and the application interface. The aim of the G3DMS is to help discover the genetic diseases prevalent in the region with the related risk factors, especially those that can be prevented. Therefore, in addition to assisting the registration of all patient data in the diagnosis process, one of the basic requirements of the system is dealing with genetic research studies and answering biological and statistical questions using relationships between conditions and family history to answer questions such as the association between maternal age and genetic disorders in the local population; and allocating cases with both genetic disorders and non-genetic diseases. The current process of answering these kinds of questions is through exhaustive manual research. Researchers and medical practitioners expect the database to be able to accommodate both clinical and genetic data, including unstructured data such as photos, patient consent forms, patients' pedigree chart, literature and patient reports. Therefore, the database should be able to provide reports and answer queries such as:

- the common genetic diseases among patients in the database
- the presence of consanguinity patients with specific genetic disorders
- the relation between mothers of advanced maternal age and specific chromosomal disorders
- group patients with common phenotypes regardless of their diagnosis
- search and compare a newly discovered mutation with the mutations in other patients
- group all cases that underwent specific diagnostic tests and compare their results.

7.3.2.2 Analysis phase

In the strategy phase, all the requirements are gathered from the case study, the basic process model and the data elements. The results identify the data flow between the processes and entities presented in the context diagram (data flow diagram level zero), the relationships between the entities as shown in the basic entity-relationship diagram, and the system requirements. In this phase, both process requirements analysis and data requirements analysis are performed to identify the best data modelling design option for the database based on the analysis outcomes. The results of this stage are the logical process model for the interface design and the conceptual data model or the global schema for the database design.

Process requirements analysis

Given the comprehensive description of the activities of individuals within the centre for the diagnosis process, the data flow is the foundation for the application interface design. First, we define the system properties that we consider necessary in building this system according to the requirements of the operations in the centre.

System specifications

- The system must provide authorised access to users (physicians/researchers), giving administrators greater privileges to manage access control.
- The system should allow the capture of patient clinical data as well as genetic results and reports.
- The system must provide a seamless, web-based user interface for end-user interaction.
- The system must allow patients' record information to be viewed, searched and modified.
- The system should answer the anticipated queries and enable researchers to build and customise their questions using the existing fields from the database, avoiding the complexity of query writing.
- The system should allow old data that may exist in the form of Excel files to be imported and integrated.
- The system should follow standardised data collection and storage procedures to enable collaboration and data sharing with other systems.
- The system should allow the export of data from the database in a useful format for integration and dataset sharing.

Based on the requirements, the G3DMS needs to accommodate patients' clinical and genetic data for care and research purposes suitable for any hereditary disease's clinic or research centre. The relational database model will suit the design of the database for this system according to the requirement for storage and data management for research. Features such as integrity constraints and normalisation will further improve the relational model by simplifying data management and data retrieval and making it possible to answer both simple and complex queries. In addition, it supports multiple users and is easy to modify without affecting the entire model body.

Logical process flow

Based on the system specifications and the context diagram (data flow diagram level zero), we prepared the logical process flow for the sub-processes and delivered the data flow diagram (level 1) presented in Figure 7.6, which outlines the primary processes performed by the physician (researcher) such as process 1.0 to add a new patient and process 2.0 to display a patient. Figure 7.6 shows the data flow between the external entity, the researcher, and each process is represented as an input and output procedure, and between the process and the patient database as a store and read operation. However, for further analysis, it is necessary to decompose the processes. Therefore, the next step is data flow diagram level 2, Figure 7.7, which goes one step deeper and shows a more detailed sub-process. For example, the process of adding a new patient to the system is subdivided into sub-processes such as 1.0 to register demographic information, 1.1 to register clinical phenotypes and 1.2 to register family history showing the data flow between the physician and the system. In every sub-process, the data flow towards the database is presented to show the data storage to keep the data which are necessary to perform other processes.

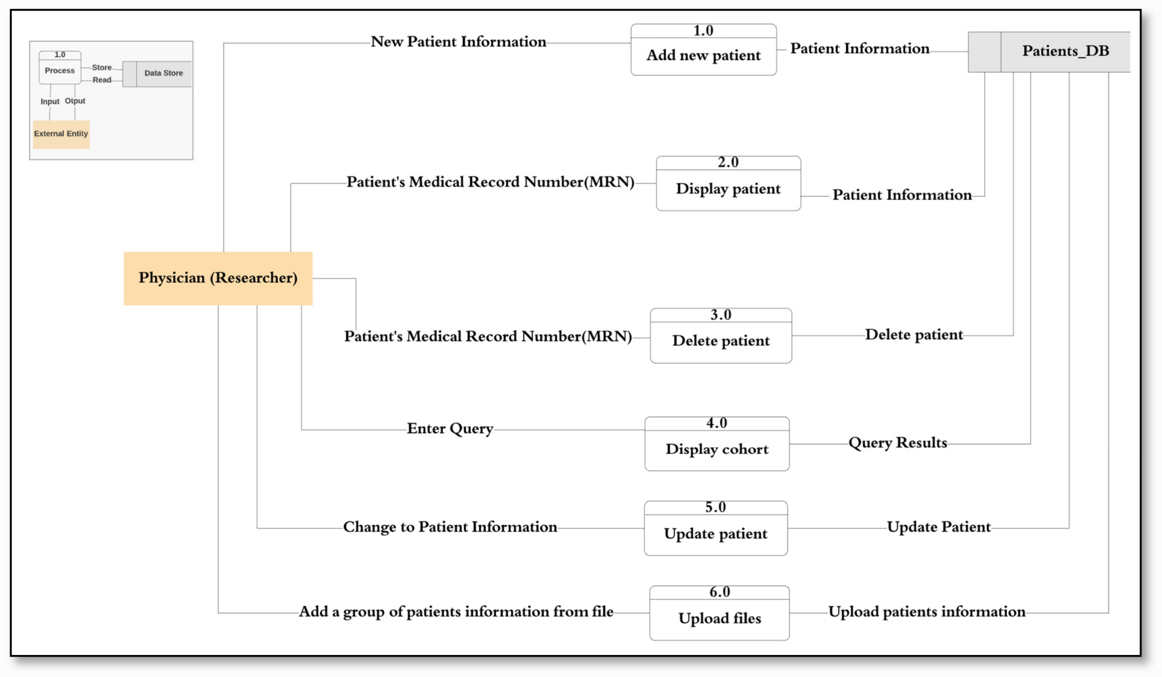


Figure 7.6: Logical process flows for data flow diagram level 1

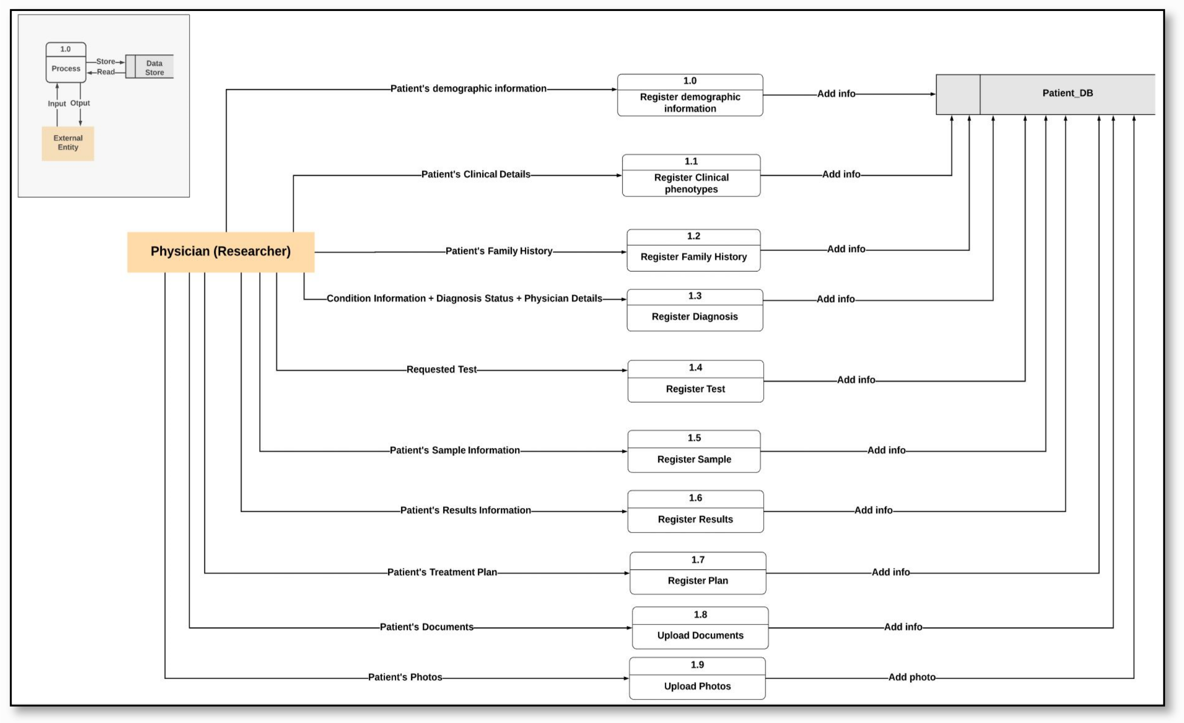


Figure 7.7: Logical process flows for data flow diagram level 2

Data requirements analysis

Based on the data elements provided in Table 7.1 from the case study of PACER-HD and the primary entity-relationship diagram in the strategy phase, we define the data elements required for PACER-HD database modelling. There is a need to group fields with shared characteristics into categories to form new entities to minimise the number of tables. For example, DEMOGRAPHICS arranges all the patient's information such as Medical Record Number (MRN), GN, name, DOB, gender, nationality and contact, and FAMILY_HISTORY hosts all the family history associated with genetic information such as Inheritance Pattern, Consanguinity, Mother's Age and DNA. There is also a need to break down some fields into more entities to increase the ability to answer more precise queries. For example, GENETIC INVESTIGATION and NON-GENETIC INVESTIGATION do not include any information about the test type, the description, the sample type or the result. New fields and entities are added to include new areas that are required to answer general research questions such as the CONDITION name and CLINICAL phenotype description in the diagnosis process. This will help with the research analysis and answer queries related to this area.

Identify entities and attributes

The research processes in the PACER-HD clinics were broken down into small elements which are mainly about the patient and diagnosis and may include other supporting elements such as genetic and non-genetic tests to achieve the relevant diagnosis. The data elements identified in Table 7.1 can be organised into entities and attributes with the relationships between them, forming a conceptual representation. The thesis follows three guidelines for classifying entities and attributes [178]:

- Entities should contain descriptive information.
- Multivalued attributes (an attribute that can have more than one value associated with the key of the entity) should be classified as entities.
- Attributes should be attached to the entities they describe.

Based on the guidelines for classification, the data elements and considering the patient-centred design, the PATIENT entity is the central entity, and all other entities have a direct relationship, such as DEMOGRAPHICS, FAMILY_HISTORY, DOCUMENT, PHOTO, CLINICAL, CONDITIONS, PHYSICIAN, TEST, SAMPLE and PLAN:

- one-to-one relationship between PATIENT and DEMOGRAPHICS and FAMILY_HISTORY entities as each patient has corresponding demographic and family history details, and each demographic and family history entity belongs to a single patient
- one-to-many relationship between PATIENT and DOCUMENT and PHOTO entities as each patient may have many documents and photos, and in return, every document or photo should belong to a single patient
- many-to-many relationship between PATIENT and entities such as CLINICAL, PHYSICIAN, CONDITIONS, SAMPLE, PLAN and TEST for example, as each patient may be diagnosed with many conditions and any condition may be detected in many patients.

Table 7.3 shows the entities and the data elements that describe the entity characteristics called attributes, in addition to the identifiers, which are the primary keys used to uniquely identify each entity. For instance, patient MRN is used as the primary key for the PATIENT.

Table 7.3: Entities and attributes type

Data Element	Description	Classification	Type	Role
PATIENT/Entity				
MRN	Patient’s Medical Record Number	Attribute	Identifier	Primary Key
DEMOGRAPHICS/Entity				
DEMOGRAPHICS_ID	Uniquely identify the entity	Attribute	Identifier	Primary Key
NAME	Patient’s name	Attribute	Descriptor	
PATIENT_GN	Patient’s Genetic Number	Attribute	Descriptor	
GENDER	Patient’s sex	Attribute	Descriptor	
DOB	Patient’s date of birth	Attribute	Descriptor	
NATIONALITY	Patient’s nationality	Attribute	Descriptor	
ADDRESS	Patient’s address	Attribute	Descriptor	
CONTACT	Patient’s mobile number	Attribute	Descriptor	
FAMILY_HISTORY/Entity				
FAMILY_HISTORY_ID	Uniquely identify the entity	Attribute	Identifier	Primary Key
INHERITANCE_PATTERN	Patient’s inheritance pattern	Attribute	Descriptor	
CONSANGUINITY	Parent’s consanguinity	Attribute	Descriptor	
MOTHER_AGE	Patient’s mother’s age	Attribute	Descriptor	
DNA_AVAILABILITY	Patient’s DNA test results	Attribute	Descriptor	
PEDIGREE_NAME	Filename for a chart tracing the inheritance of one or more traits through a family	Attribute	Descriptor	
PEDIGREE_TYPE	File type	Attribute	Descriptor	
PEDIGREE_CONTENT	Data (content)	Attribute	Descriptor	
PHYSICIAN/Entity				
PHYSICIAN_ID	Uniquely identify the entity	Attribute	Identifier	Primary Key
PHYSICIAN _NAME	Physician’s name	Attribute	Descriptor	
CONTACT	Physician’s mobile number	Attribute	Descriptor	
CONDITIONS/Entity				
CONDITION_NAME	Patient’s condition’s name	Attribute	Identifier	Primary Key
CLINICAL/Entity				
CLINICAL_ID	Uniquely identify the entity	Attribute	Identifier	Primary Key
CLINICAL_DESCRIPTION	Phenotypes information of the patient which is related to the specific condition	Attribute	Descriptor	
TEST/Entity				
TEST_ID	Uniquely identify the entity	Attribute	Identifier	Primary Key
TEST_TYPE	Specific test type	Attribute	Descriptor	
SAMPLE/Entity				
SAMPLE_ID	Uniquely identify the entity	Attribute	Identifier	Primary Key
SAMPLE_TYPE_	Specify the sample type	Attribute	Descriptor	
PLAN/Entity				
PLAN_ID	Uniquely identify the entity	Attribute	Identifier	Primary Key
PLAN_TYPE	Name of the plan	Attribute	Descriptor	
PHOTO/Entity				
PHOTO_ID	Uniquely identify the entity	Attribute	Identifier	Primary Key
PHOTO _NAME	Name of the photo	Attribute	Descriptor	
PHOTO _TYPE	Type of photo	Attribute	Descriptor	
PHOTO _CONTENT	Patient’s photo (data)	Attribute	Descriptor	
DOCUMENT/Entity				
DOCUMENT _ID	Uniquely identify the entity	Attribute	Identifier	Primary Key
DOC_NAME	File name	Attribute	Descriptor	
DOC_TYPE	File Type	Attribute	Descriptor	
DOC_CONTENT	Any important notes, reports, and literature in PDF format (data)	Attribute	Descriptor	
DOC_DESCRIPTION	Text to describe the document type	Attribute	Descriptor	

Global schema

The entity-relationship diagram is used to satisfy the data modelling objectives and develop the conceptual model or the global schema, which is used to construct the physical database structure. As shown in Figure 7.8, the diagram outlines all the details of entities with their attributes defined in the previous section as well as identifiers and constraints. In the diagram, the primary keys are underlined to distinguish them from the attributes of the other descriptors. The cardinality was specified by minimum and maximum values (one, many, one or many, and zero or many). This is to show the constraints, or a restriction placed on the data to ensure data integrity. For example, for a patient to be included in this database, there should be at least one condition (one or many); however, a patient may or may not have documents (zero or many). The diagram also shows two types of relationships.

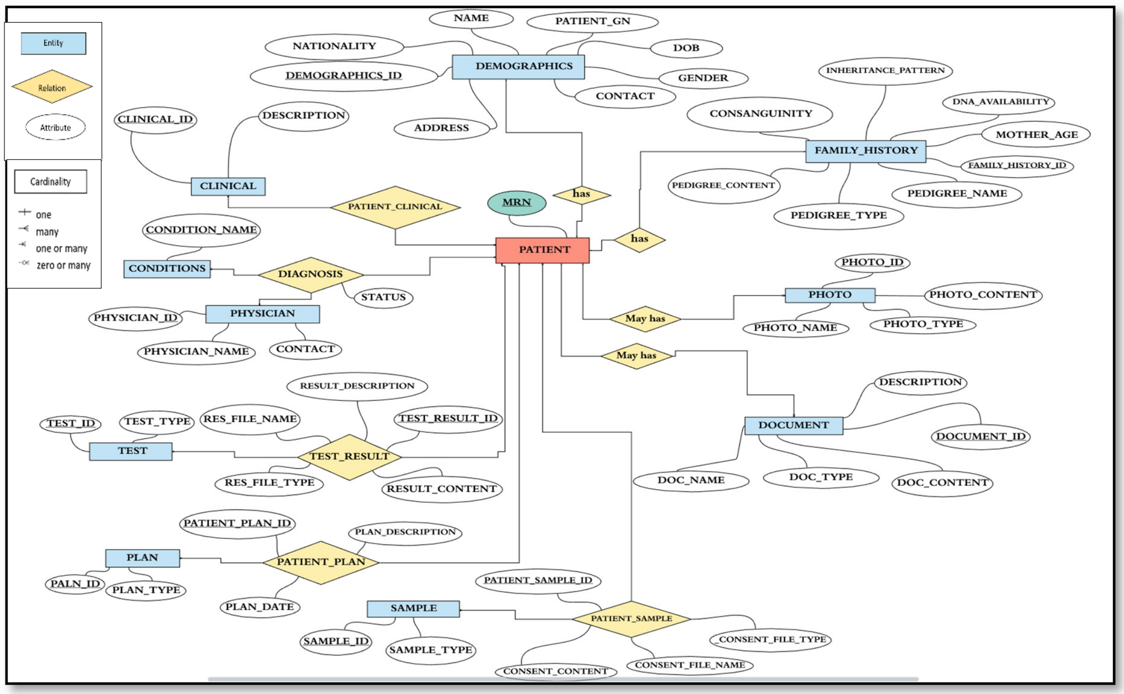


Figure 7.8: Entity–relationship diagram (the global schema)

7.3.2.3 Design phase

The outputs of the analysis stage are the logical process flow which includes the data flow diagrams levels 1 and 2 as well as the conceptual data model or the global schema. This phase aims to develop the logical data model or logical schema from the conceptual model and then convert the logical schema to the physical data model, i.e. constructs the database tables and references. It proceeds towards the design of the application interface depending on the logical process models and prepares the physical process flow and the interface wireframe diagrams.

Logical data model

Throughout this phase, the focus is on developing the physical model by converting the logical schema that has been built based on the entity-relationship diagram for data requirements specification and the conceptual model defined in the previous two-phase strategy and analysis. The data model described in the diagram in Figure 7.8 is converted to suit the database management system, so entities, attributes and relationships must be converted into tables and fields. Therefore, any entity will be converted to a table, its attributes will become fields, and its primary key will become the primary key of the table. In the case of the many-to-many relationship between two entities (binary) such as PATIENT and SAMPLE, a third table will be created with the name PATIENT_SAMPLE. Similarly, in the case of many-to-many between three entities (turnery), for example, PATIENT, PHYSICIAN and CONDITIONS, a third table will be created with the name DIAGNOSIS by combining the three primary keys from the participating entities. The normalisation technique is applied to refine the design to remove duplication (uncontrolled data redundancy) that may occur in tables to eliminate the problem of modification anomalies (while adding, deleting and modifying records in a table). The logical structure of the database (schema), namely the structural representation of what is in the database is presented in Figure 7.9, which clearly defines the tables (entities), the fields (attributes) in the tables, the primary and foreign keys, and the relationships between tables.

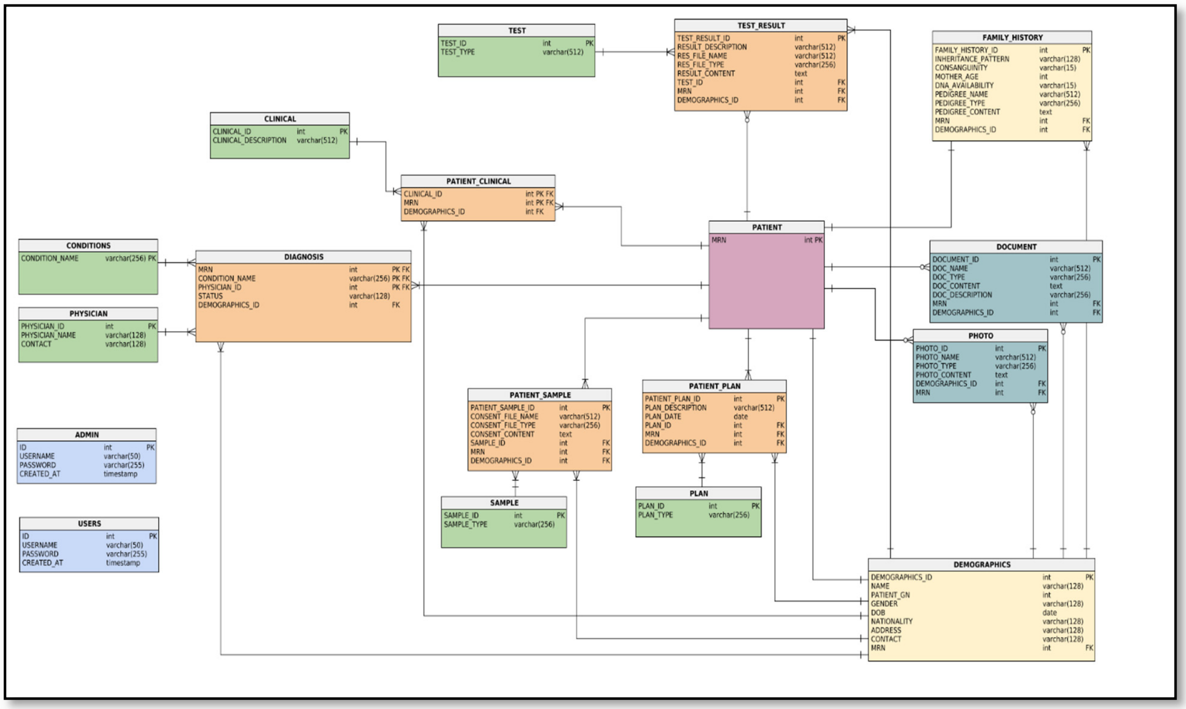


Figure 7.9: Logical schema

Physical data model

The logical schema in Figure 7.9 is converted to a physical model where entities are mapped into tables, instances into rows, and attributes into columns. Each table contains the column names and its specified data type, and the primary keys are also determined for each table which can be a single or composite key. References are used to describe the relationships between tables recognised by foreign keys; references facilitate the indexing mechanism for efficient access to the data. Figure 7.10 shows the physical representation of a many-to-many relationship as tables, for example, the DIAGNOSIS table with its composite primary key (MRN, CONDITION_NAME and PHYSICIAN_ID) and their respective tables. However, the linkage between the entities of a one-to-many relationship can be represented using a referencing key or foreign key. The data dictionary for G3DMS is in Appendix 7.1.

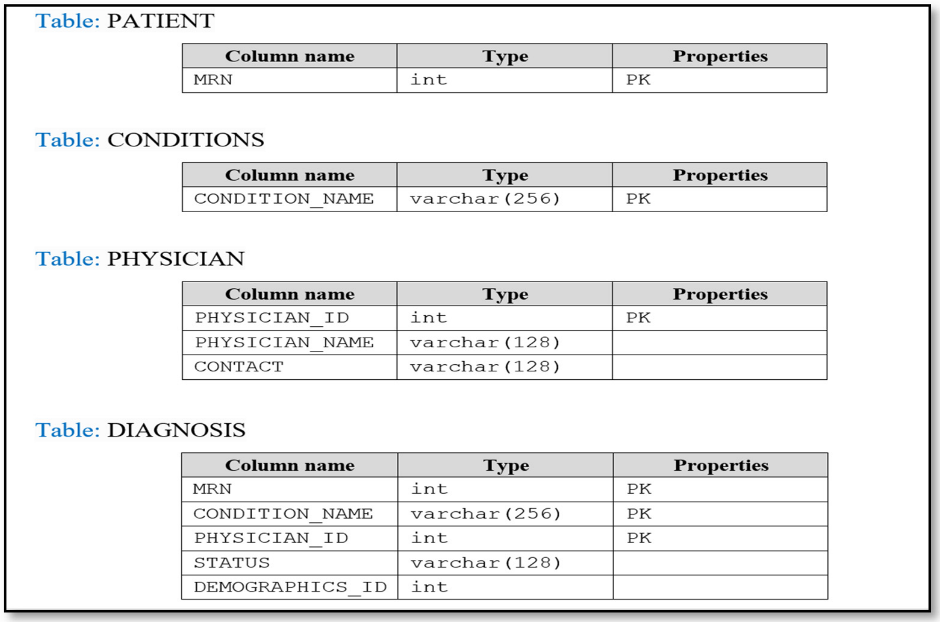


Table: PATIENT

Column name	Type	Properties
MRN	int	PK

Table: CONDITIONS

Column name	Type	Properties
CONDITION_NAME	varchar(256)	PK

Table: PHYSICIAN

Column name	Type	Properties
PHYSICIAN_ID	int	PK
PHYSICIAN_NAME	varchar(128)	
CONTACT	varchar(128)	

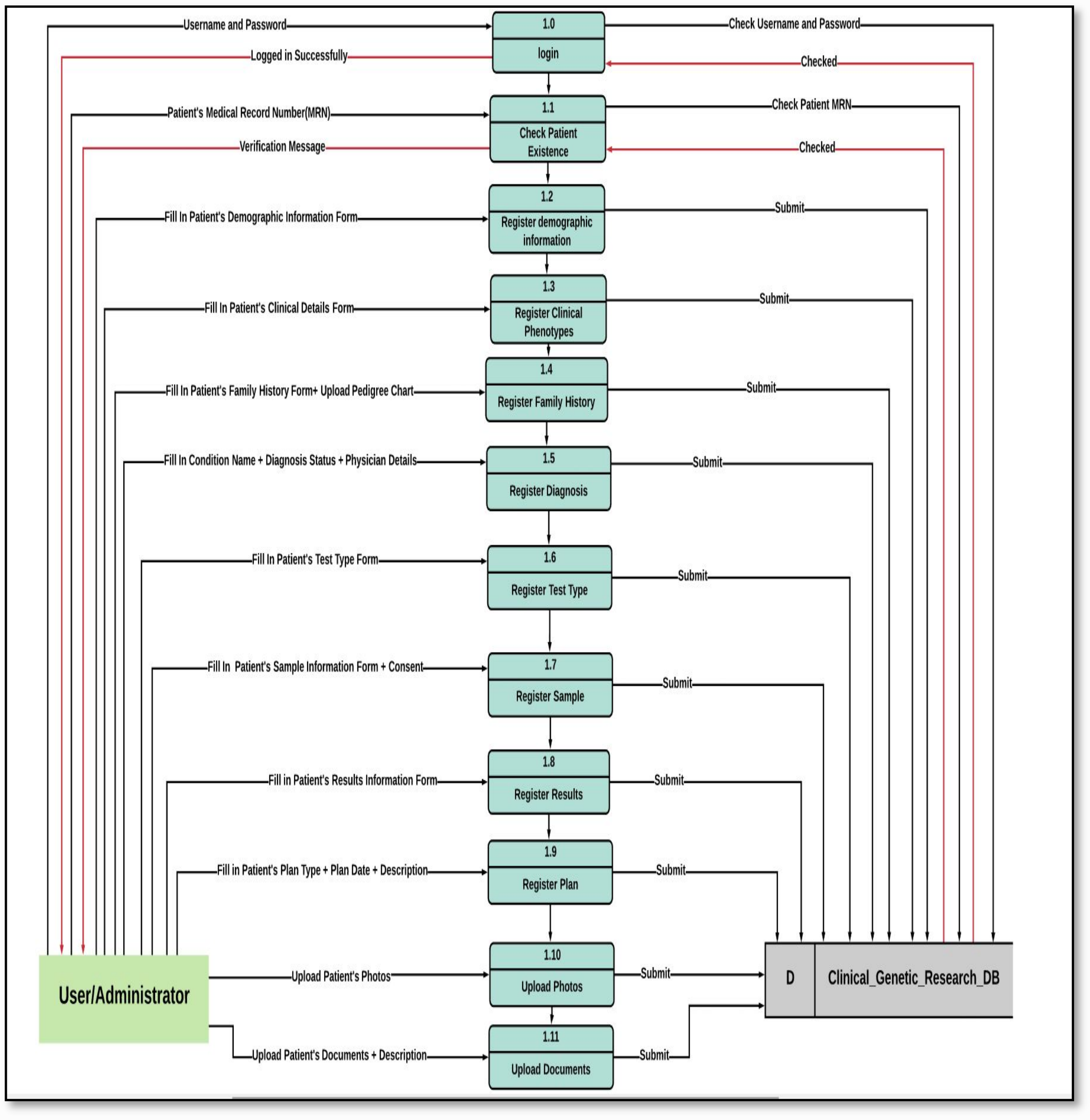
Table: DIAGNOSIS

Column name	Type	Properties
MRN	int	PK
CONDITION_NAME	varchar(256)	PK
PHYSICIAN_ID	int	PK
STATUS	varchar(128)	
DEMOGRAPHICS_ID	int	

Figure 7.10: Converting the logical schema to physical model and tables

Physical process flow

In addition to the logical database schema, the product of this stage is the physical process flow in Figure 7.11 which is based on the logical process flow in the analysis phase. The focus of this model is on the detailed description of the system and user interactions. The physical process flow determines how the user interface will be designed and how forms are constructed to collect data from the user and submit it to the database. For example, the process of adding a new patient is logically illustrated using the data flow diagrams. The physical process flow shows a comprehensive view from the perspective of the application interface interaction, including all responses in the form of messages to each user action.



Username and Password
Logged in Successfully
1.0
login
Check Username and Password
Checked
Patient's Medical Record Number(MRN)
Verification Message
1.1
Check Patient Existence
Check Patient MRN
Checked
Fill In Patient's Demographic Information Form
1.2
Register demographic information
Submit
Fill In Patient's Clinical Details Form
1.3
Register Clinical Phenotypes
Submit
Fill In Patient's Family History Form+ Upload Pedigree Chart
1.4
Register Family History
Submit
Fill In Condition Name + Diagnosis Status + Physician Details
1.5
Register Diagnosis
Submit
Fill In Patient's Test Type Form
1.6
Register Test Type
Submit
Fill In Patient's Sample Information Form + Consent
1.7
Register Sample
Submit
Fill in Patient's Results Information Form
1.8
Register Results
Submit
Fill in Patient's Plan Type + Plan Date + Description
1.9
Register Plan
Submit
User/Administrator
Upload Patient's Photos
1.10
Upload Photos
Submit
D
Clinical_Genetic_Research_DB
Upload Patient's Documents + Description
1.11
Upload Documents
Submit

Figure 7.11: Physical process flow for adding new patient

Interface design

The process flow models are translated into the actual structure of the application interface, which will be developed accordingly. The site map in Figure 7.12 illustrates the plan for the organisation of the web pages according to the intended user action to perform a specific task within any process flow. This hierarchical representation clearly shows the parent-child relationship between the pages of the website. The home page provides an authorised entry to the system according to the role of user or administrator. The admin web page provides access to pages to reset passwords and register new users or admin, as access to this system should be approved by the appropriate organisational authority. Users can also maintain their account and have access to the reset password page. Users who have both roles can access the G3DMS main page, which can provide access to four pages: Add New Patient, Query Builder, Upload Data File, and Export the Database. This page allows the user to move sequentially to the next page to fill in all the patient data using text boxes, starting with the demographic information page to the plan, documents and photos page. The query builder page provides access to four pages to display the selected patient's information, display a cohort with common characteristics and research queries, delete a patient, and update patient information.

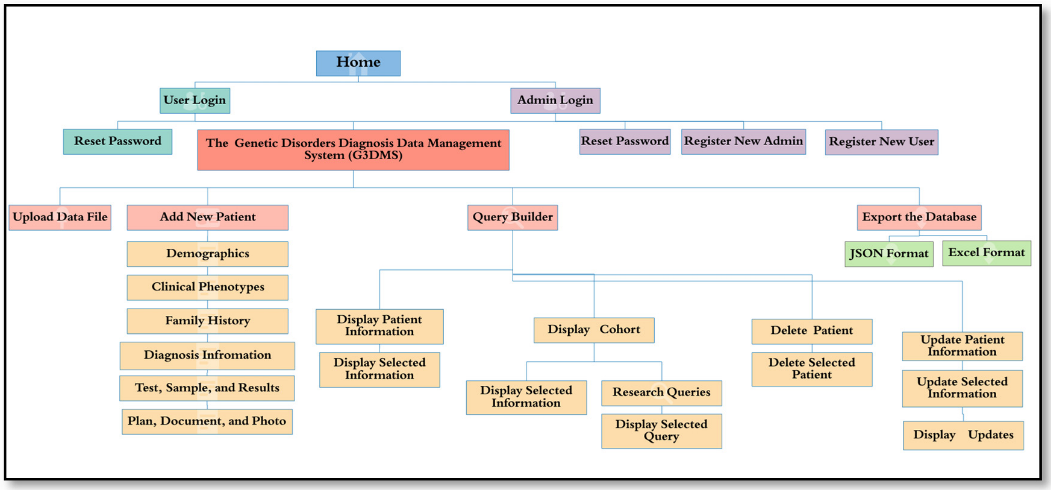
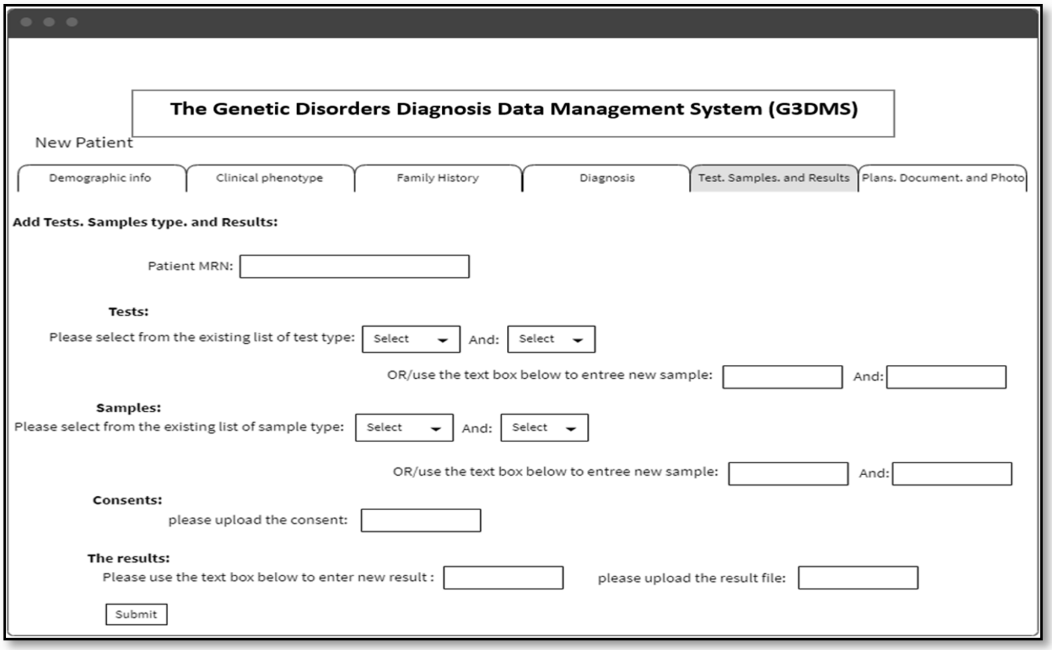


Figure 7.12: Site map of the G3DMS

Wireframe diagrams

While the site map shows the organisational and hierarchical presentation of the web pages, wireframe diagrams outline the layout and element functionality within the page and the site navigation details. Figure 7.13 presents a sample page of the interface using the wireframe diagram. An example of adding new patient data is provided in the layout; the page for filling in the diagnostic tests requested by the physician, the patient sample data with the consent document, and the tests requested with its results can be in the form of text or uploaded as a file. Full site wireframe diagrams are available in Appendix 7.2.



The Genetic Disorders Diagnosis Data Management System (G3DMS)

New Patient

Add Tests. Samples type. and Results:

Patient MRN:

Tests:

Please select from the existing list of test type: Select And: Select

OR/use the text box below to entree new sample: And:

Samples:

Please select from the existing list of sample type: Select And: Select

OR/use the text box below to entree new sample: And:

Consents:

please upload the consent:

The results:

Please use the text box below to enter new result : please upload the result file:

Figure 7.13: Wireframe diagram for adding new patient: tests, sample and results page

7.4 Summary

This chapter provided the three fundamental stages in the systems development lifecycle of the G3DMS. The strategy phase presented the actual problem using a case study from the Saudi context of the PACER-HD centre to show the limitation with the current system. The strategy stage resulted in three fundamentals requirements: the basic entity-relationship diagram for the required data elements, the data flow diagram for the process flow, and the future system requirements. In the analysis phase, the requirements from the previous stage were analysed to produce the logical process flow in the form of data flow diagrams and the conceptual data model which was represented in the entity-relationship diagram and the global schema. Next, the results of the process and data models were used in the design phase to develop the physical models for the application interface and the database, respectively. Next, the Barker systems development lifecycle for the G3DMS presented in Figure 7.2 proceeds in Chapter 8 with the remainder of the development and implementation phases: build, documentation, transition and production.

Chapter 8: Genetic Disorders Diagnosis Data Management System (G3DMS): Build, Documentation, Transition and Production

8.0 Chapter overview

This chapter illustrates the implementation process of the novel genetic disorders diagnosis data management system (G3DMS) and contains the next four phases of the systems development lifecycle with the Barker method presented in Figure 7.3. Section 8.1 presents the aims and purpose of this chapter and provides connections to previous chapters. Section 8.2 summarises the design methodology of the first three stages of the Barker method in Chapter 7 and proceeds with the Barker method systems development lifecycle stages: build, documentation, transition and production. Section 8.3 presents qualitative evaluation methods for the system design and implementation of the G3DMS. Section 8.4 shows the results of the proposed data management system, the G3DMS. Section 8.5 highlights the contribution and future work of the G3DMS. Finally, Section 8.6 summarises the chapter and introduces the next chapter.

8.1 Introduction

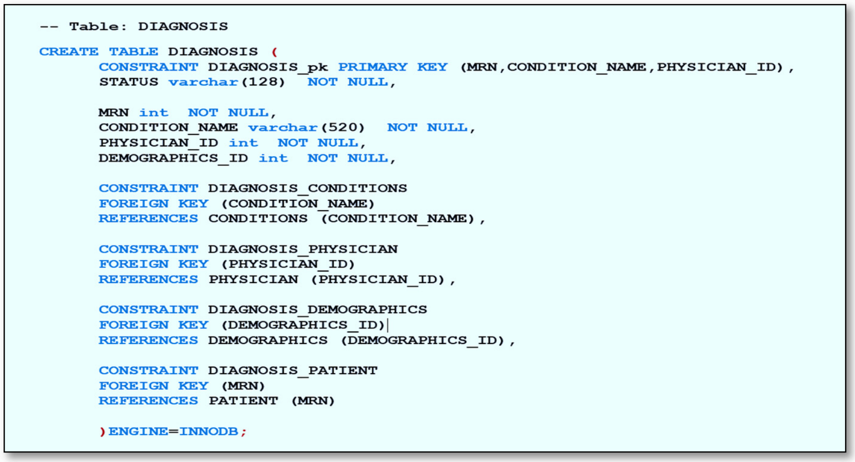
This chapter presents the last step in the implementation stage and delivers the G3DMS to the real world (Figure 1.1, Figure 1.2). The purpose of this chapter is to produce the finished product of the G3DMS, and answer the research question, Figure 1.1, Q5.B: *How can a data management system be implemented?* As the previous chapter delivered the first three stages of the systems development lifecycle, this chapter proceeds with the Barker method and delivers the next four stages of build, documentation, transition and production, as well as using qualitative evaluation methods for successful health information system implementation.

8.2 Build phase

Building the database involves the actual creation of the whole system, the application interface, and the database that has been physically designed in the design phase. In this phase, the programming code and the database code are written. The build phase takes place in the chosen test environment to build the initial schema before the implementation in the production environment. First, the choice to build the database using the relational model as specified in the analysis and the design phase according to the requirement analysis and the purpose of the G3DMS facilitates the data structure and management using the right relational database management system for such a system. Therefore, the creation of the database depends on the selection of the storage container or server and the database management system to run the database scripts. Second, the application interface development option will also affect the entire system performance; therefore, selecting the environment that supports the structured query language as well as the interface programming is very crucial. Thus, in this phase, the aim is to choose the development environment that accommodates the design for the backend database and the frontend application interface, including the user interface.

8.2.1 Build the database

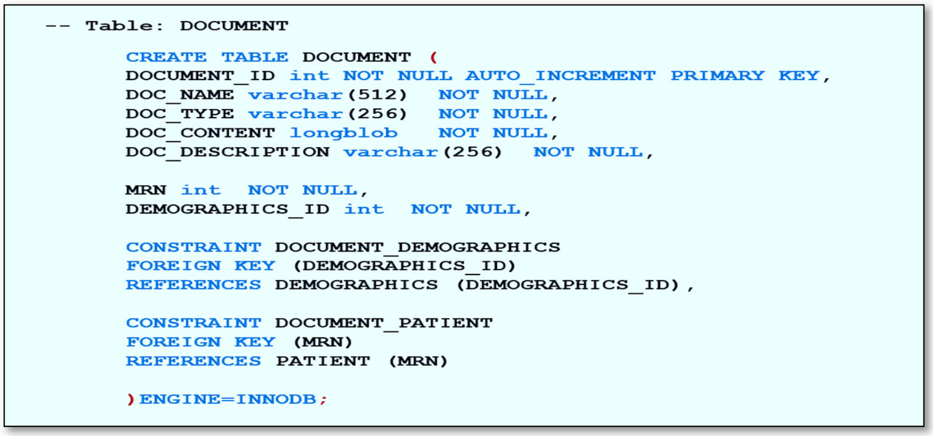
For the database construction, the physical data models (tables and references) are implemented in phpMyAdmin server. First, the database is created followed by the PATIENT and the DEMOGRAPHICS tables, on which all the other tables in the database depend. After this, the independent tables are created, such as CLINICAL, CONDITIONS, PLAN, SAMPLE, TEST and PHYSICIAN, which store a list of shared information that may be linked to any patient in the database. Then, the patient tables that depend on the previous tables are created, such as DIAGNOSIS which depends on PATIENT, CONDITIONS and PHYSICIAN, in addition to DEMOGRAPHICS for easy access to the patient's personal information. Figure 8.1 shows the script for constructing the table DIAGNOSIS showing the primary key and all its attributes as well as the foreign keys that link the table to its associates. Tables such as FAMILY_HISTORY, DOCUMENT and PHOTO, which are directly dependent on the PATIENT table are also created similarly.



```sql
-- Table: DIAGNOSIS

CREATE TABLE DIAGNOSIS (
        CONSTRAINT DIAGNOSIS_pk PRIMARY KEY (MRN,CONDITION_NAME,PHYSICIAN_ID),
        STATUS varchar(128)  NOT NULL,

        MRN int  NOT NULL,
        CONDITION_NAME varchar(520)  NOT NULL,
        PHYSICIAN_ID int  NOT NULL,
        DEMOGRAPHICS_ID int  NOT NULL,

        CONSTRAINT DIAGNOSIS_CONDITIONS
        FOREIGN KEY (CONDITION_NAME)
        REFERENCES CONDITIONS (CONDITION_NAME),

        CONSTRAINT DIAGNOSIS_PHYSICIAN
        FOREIGN KEY (PHYSICIAN_ID)
        REFERENCES PHYSICIAN (PHYSICIAN_ID),

        CONSTRAINT DIAGNOSIS_DEMOGRAPHICS
        FOREIGN KEY (DEMOGRAPHICS_ID)
        REFERENCES DEMOGRAPHICS (DEMOGRAPHICS_ID),

        CONSTRAINT DIAGNOSIS_PATIENT
        FOREIGN KEY (MRN)
        REFERENCES PATIENT (MRN)

        )ENGINE=INNODB;
```

Figure 8.1: SQL script for creating the DIAGNOSIS table

Data types are stored as specified in the database library in the physical data model in Appendix 7.1, except for the document contents, and the photos are stored as a binary large object (BLOB) data type. Storing patient data files such as documents, reports and photos in a long binary large object (LONGBLOB) format allows us to keep the content in the database and retrieve it without any damage. Figure 8.2 shows the script used to define a document using a name, type and content in the DOCUMENT table. Figure 8.3 shows the complete database in phpMyAdmin interface using the localhost. The full SQL scripts for the database construction are in Appendix 8.1.



```
-- Table: DOCUMENT

      CREATE TABLE DOCUMENT (
      DOCUMENT_ID int NOT NULL AUTO_INCREMENT PRIMARY KEY,
      DOC_NAME varchar(512)  NOT NULL,
      DOC_TYPE varchar(256)  NOT NULL,
      DOC_CONTENT longblob   NOT NULL,
      DOC_DESCRIPTION varchar(256)  NOT NULL,

      MRN int  NOT NULL,
      DEMOGRAPHICS_ID int  NOT NULL,

      CONSTRAINT DOCUMENT_DEMOGRAPHICS
      FOREIGN KEY (DEMOGRAPHICS_ID)
      REFERENCES DEMOGRAPHICS (DEMOGRAPHICS_ID),

      CONSTRAINT DOCUMENT_PATIENT
      FOREIGN KEY (MRN)
      REFERENCES PATIENT (MRN)

      )ENGINE=INNODB;
```

Figure 8.2: SQL script for using LONGBLOB data type in the DOCUMENT table

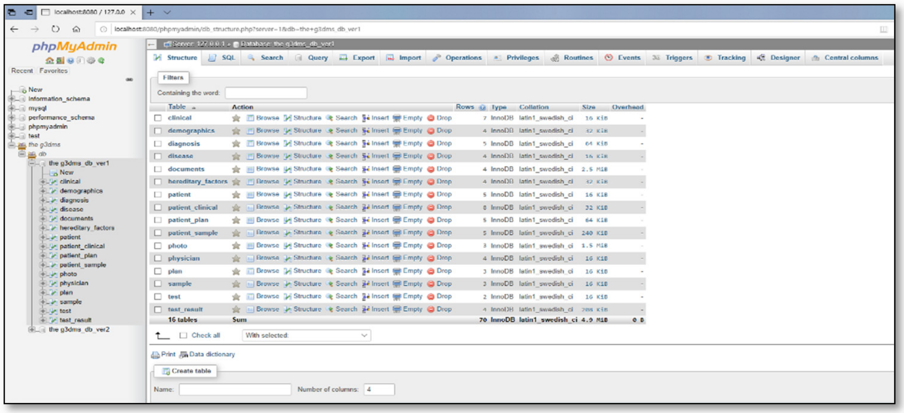


	Table	Action						Rows	Type	Collation	Size	Overhead
□	clinical	☆	Browse	Structure	Search	Insert	Empty	Drop	7	InnoDB	latin1_swedish_ci	16 KiB	-
□	demographics	☆	Browse	Structure	Search	Insert	Empty	Drop	4	InnoDB	latin1_swedish_ci	32 KiB	-
□	diagnosis	☆	Browse	Structure	Search	Insert	Empty	Drop	5	InnoDB	latin1_swedish_ci	64 KiB	-
□	disease	☆	Browse	Structure	Search	Insert	Empty	Drop	4	InnoDB	latin1_swedish_ci	16 KiB	-
□	documents	☆	Browse	Structure	Search	Insert	Empty	Drop	4	InnoDB	latin1_swedish_ci	2.5 MiB	-
□	hereditary_factors	☆	Browse	Structure	Search	Insert	Empty	Drop	4	InnoDB	latin1_swedish_ci	32 KiB	-
□	patient	☆	Browse	Structure	Search	Insert	Empty	Drop	5	InnoDB	latin1_swedish_ci	16 KiB	-
□	patient_clinical	☆	Browse	Structure	Search	Insert	Empty	Drop	8	InnoDB	latin1_swedish_ci	32 KiB	-
□	patient_plan	☆	Browse	Structure	Search	Insert	Empty	Drop	5	InnoDB	latin1_swedish_ci	64 KiB	-
□	patient_sample	☆	Browse	Structure	Search	Insert	Empty	Drop	5	InnoDB	latin1_swedish_ci	240 KiB	-
□	photo	☆	Browse	Structure	Search	Insert	Empty	Drop	3	InnoDB	latin1_swedish_ci	1.5 MiB	-
□	physician	☆	Browse	Structure	Search	Insert	Empty	Drop	4	InnoDB	latin1_swedish_ci	16 KiB	-
□	plan	☆	Browse	Structure	Search	Insert	Empty	Drop	3	InnoDB	latin1_swedish_ci	16 KiB	-
□	sample	☆	Browse	Structure	Search	Insert	Empty	Drop	3	InnoDB	latin1_swedish_ci	16 KiB	-
□	test	☆	Browse	Structure	Search	Insert	Empty	Drop	2	InnoDB	latin1_swedish_ci	16 KiB	-
□	test_result	☆	Browse	Structure	Search	Insert	Empty	Drop	4	InnoDB	latin1_swedish_ci	208 KiB	-
	16 tables	Sum							70	InnoDB	latin1_swedish_ci	4.9 MiB	0 B

Figure 8.3: G3DMS database

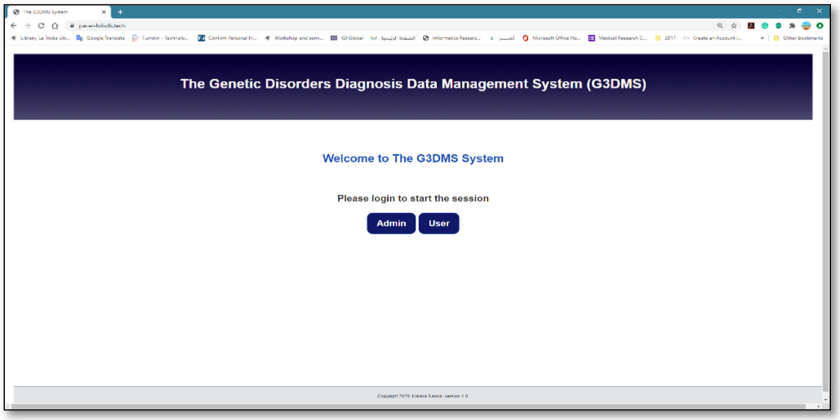
8.2.2 Build the application interface

Once the database has been created, the G3DMS application interface is ready to be constructed, considering the user interface layout based on the wireframe diagrams and linkage methods to communicate between the user interface page requests and the database response. Each wireframe design for a specific page is converted to the actual web page using HTML code to create forms to collect data using menus for selection as well as text boxes to insert data. The PHP MySQLi functions allow for connecting and communicating to the MariaDB server using MySQLi driver methods for adding, reading, updating, modifying and deleting the content of the database. An open-source, cross-platform web server solution stack package (XAMPP) platform is also used as a development environment for the frontend using the Apache server as localhost to run the source code for the web pages during the development process [207]. The complete source code for the web pages of the application interface is provided in Appendix 8.2.

Web interface pages

Login pages

Access to the system is restricted to authorised users only. Therefore, the first page in the interface in Figure 8.4 is designed to display the login options according to the user's role and their level of authorisation. The system allows logging on as a regular user or as an administrator. Administrators have additional privileges, such as being able to register new users and define their roles. Both users of the system can reset their passwords through the reset password page after logging in.



The Genetic Disorders Diagnosis Data Management System (G3DMS)

Welcome to The G3DMS System

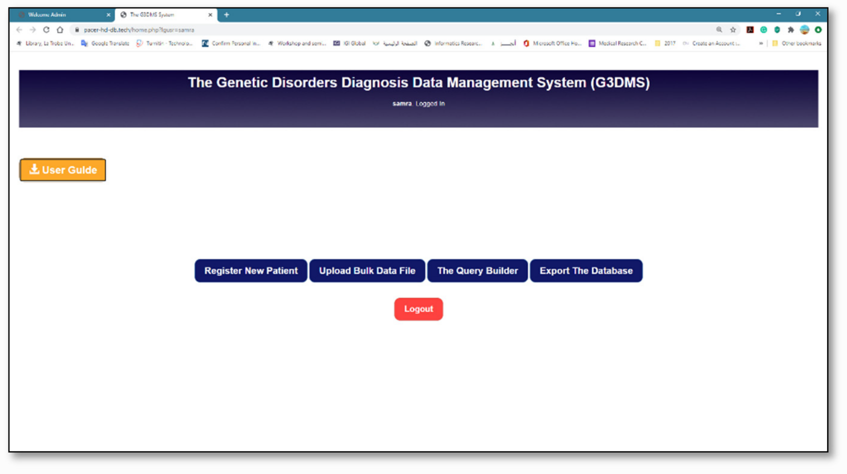
Please login to start the session

Copyright 2020; Halima Samra; version 1.0

Figure 8.4: Login page

G3DMS home page

The G3DMS home page is an HTML coded page that provides links to the Register New Patient page, the Upload Bulk Data File page, the Query Builder page and the Export the Database page. It also contains a link to download the user guide in a PDF file for more instructions and the user interaction documentation.

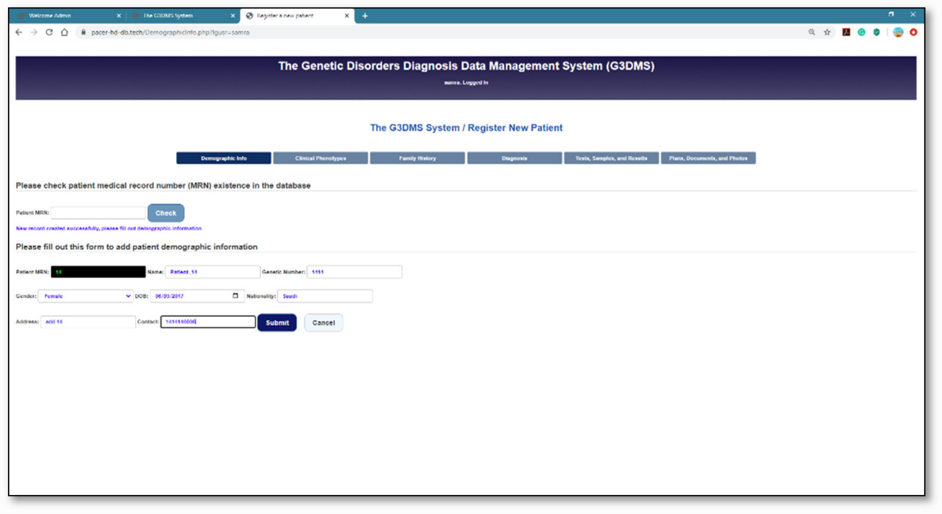

The Genetic Disorders Diagnosis Data Management System (G3DMS)

samra Logged In

Figure 8.5: G3DMS home page

Register New Patient page

The process of registering a new patient is completed in six steps to facilitate the data entry process and allow the user to focus on each step individually and prevent information clutter. The page structure starts with demographic information, as shown in Figure 8.6, followed by clinical phenotypes, family history, diagnosis information, tests, sample type, test results, plans, documents and photos. The system will allow the user to complete the patient's data sequentially; each step of the registration process is developed in an individual PHP page with HTML codes to display the forms. First, before entering the data, the system allows the patient MRN to be checked in the database to prevent data duplication and override. The MRN is checked in the PATIENT table, and then a response message is shown if it exists, but if it does not exist then a new record is created in the PATIENT table, and by submitting the form the data is sent to the other designated patient's table.


The Genetic Disorders Diagnosis Data Management System (G3DMS)

samra. Logged In

The G3DMS System / Register New Patient

Please check patient medical record number (MRN) existence in the database

Patient MRN:

New record created successfully, please fill out demographic information

Please fill out this form to add patient demographic information

Patient MRN: 14 Name: Patient_14 Genetic Number: 1414

Gender: Female DOB: 06/03/2017 Nationality: Saudi

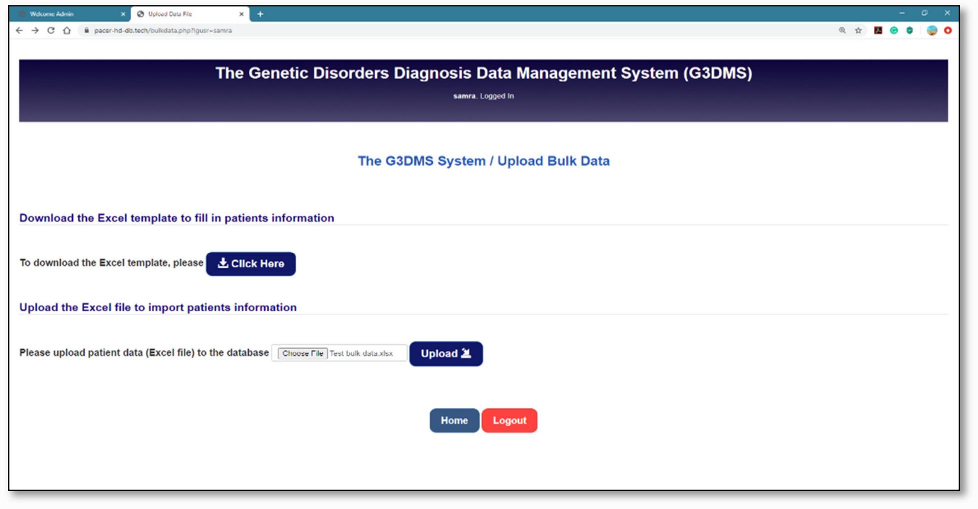
Address: add 14 Contact: 1414140000

pacer-hd-db.tech/DemographicInfo.php?lgusr=samra
Demographic Info
Clinical Phenotypes
Family History
Diagnosis
Tests, Samples, and Results
Plans, Documents, and Photos
Check
Submit
Cancel

Figure 8.6: Register New Patient page and demographic information

Upload Bulk Data File page

The G3DMS offers another method of data entry, especially for migrating old data or data collected offline for research. Figure 8.7 presents the bulk data upload page which contains two options: download the Excel template to fill in the patients' data offline, and the upload button to import bulk data using the Excel template.



The Genetic Disorders Diagnosis Data Management System (G3DMS)

samra, Logged In

The G3DMS System / Upload Bulk Data

Download the Excel template to fill in patients information

To download the Excel template, please

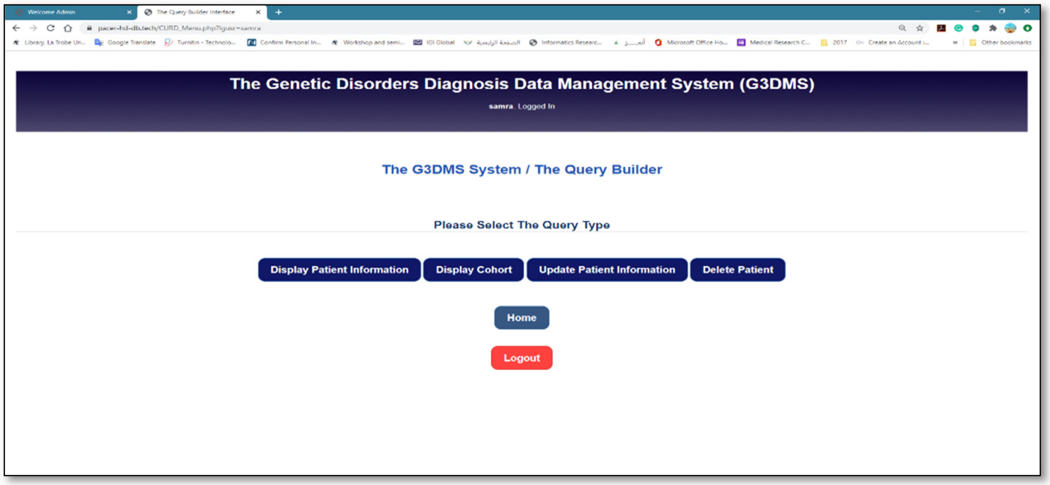
Upload the Excel file to import patients information

Please upload patient data (Excel file) to the database Test bulk data.xlsx

Figure 8.7: Upload Bulk Data download template and upload data page

Query Builder page

This page is the main access page for simple, complex and advanced queries. Figure 8.8 shows the contents of the page, which provides links to the following pages: Display Patient Information page, Display Cohort pages, Update Patient Information page and Delete Patient page. Each option links to the page that performs the selected task. The Display Cohort page provides access to perform more advanced queries through the Display Research Queries page. Additional detail about each page is provided in the user guide in Appendix 8.3.



The Genetic Disorders Diagnosis Data Management System (G3DMS)

samra, Logged In

The G3DMS System / The Query Builder

Please Select The Query Type

Figure 8.8: Query Builder main page

Export the database page

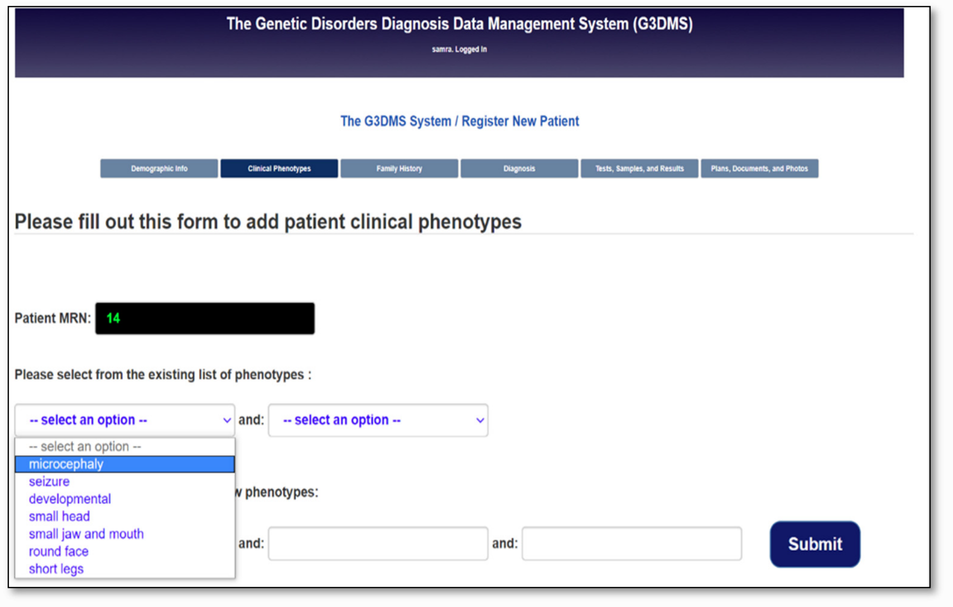
This page allows users to obtain datasets from the database in the form of Excel files and JSON files. The page provides two options: Export to Excel and Export to JSON buttons. By clicking on the button, the data are downloaded to the user's computer in the requested format.

8.2.3 Building features

The G3DMS aims to deliver an easy-to-use user interface equipped with features to facilitate both patient care and to produce quality datasets for research. Therefore, several aspects are considered while building the application interface to ensure ease of use, data quality, research results accuracy and interactive nature.

Progressive lists

As the users enter new data for the first time using text boxes, the self-developing dynamic lists are automatically updated and available to be used for data entry. The main advantage of this feature is to reduce data duplication, which affects data quality, hence playing an important role in the accuracy of the research results. It is less time consuming for users to enter data, so they can focus more on the care process instead of data entry. The progressive lists combine common characteristics such as phenotypes, conditions, physicians, sample types, tests and plan types which can be shared by multiple patients. Figure 8.9 shows an example of a progressive list for patients' phenotypes which is displayed when registering a new patient. This menu is filled in by inserting new phenotypes using text boxes which constitutes an alternative method of data entry. Then the submitted data is checked by PHP MySQLi functions, and if the inserted value does not exist, it is inserted in the CLINICAL table and referenced in the PATIENT_CLINICAL table using the ID. However, if the value exists in the table, then the CLINICAL_ID is retrieved and linked to the PATIENT_CLINICAL table. For multiple phenotypes, the entry is stored in an array and is then retrieved using FOREACH to access all the values. To display the list, as shown in Figure 8.9, a PHP code is used to read the table using the SELECT statement and then display the results inside the HTML <select> as an <option>.



The Genetic Disorders Diagnosis Data Management System (G3DMS)

samra. Logged In

The G3DMS System / Register New Patient

Please fill out this form to add patient clinical phenotypes

Patient MRN: 14

Please select from the existing list of phenotypes :

-- select an option -- and: -- select an option --

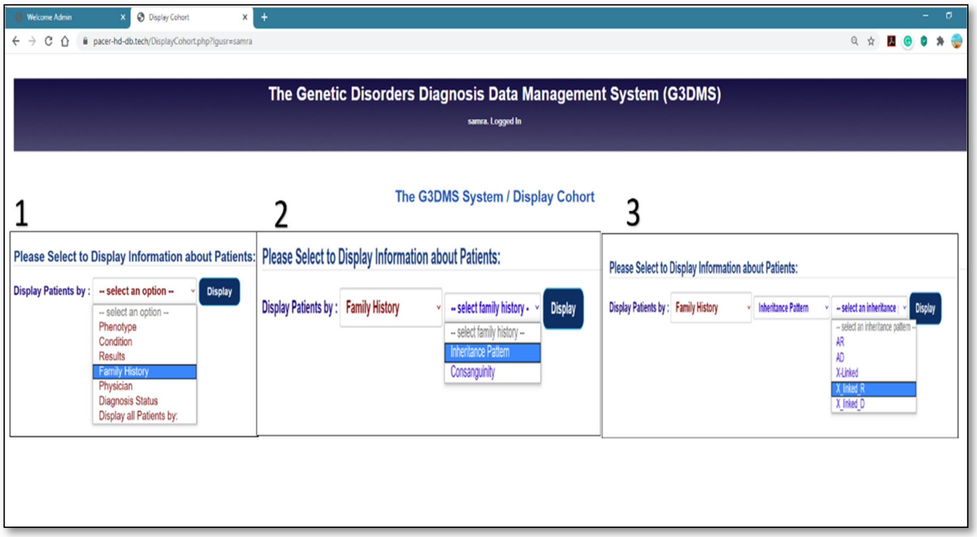
w phenotypes:

and: and:

Figure 8.9: Progressive lists example: clinical phenotype menus

Interactive menus

The page structure of the G3DMS development should be clear and avoid ambiguity. Therefore, we focus on reducing the number of unnecessary elements not related to the task at hand. To do this, we used the jQuery function to manipulate the appearance of menus on pages specifically designed to answer research questions. First, we added the jQuery function as <script> with the HTML code. Next, we defined each select list as a variable inside the function, and then used the hide () function to hide all the lists except those required by the user. This feature is implemented in the Display Cohort page which displays a group of patients with common characteristics; first, the user selects from the first dropdown list to choose the main category (phenotype, condition, results). Then, according to the user's selection in the first list, the system displays a dynamic list for that option. Figure 8.10 depicts an example of a user selecting Family History from the first menu, at which point another menu is displayed for the options (inheritance patterns or consanguinity) and depending on the user's selection, a third menu appears which displays a dynamic list of content from the database.


pacer-hd-db.tech/DisplayCohort.php?lgusr=samra

The Genetic Disorders Diagnosis Data Management System (G3DMS)

samra. Logged In

The G3DMS System / Display Cohort

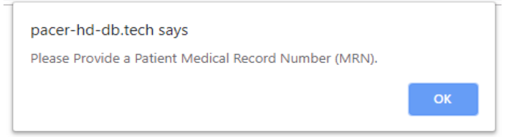
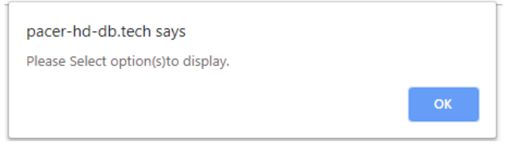
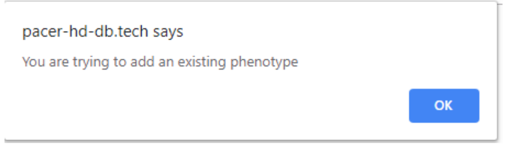
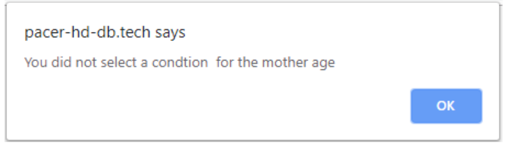
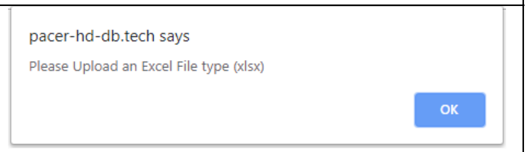
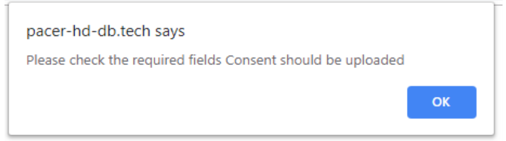
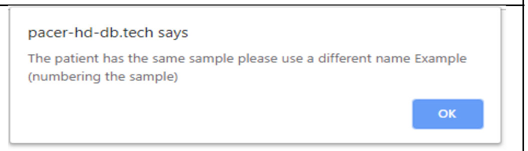
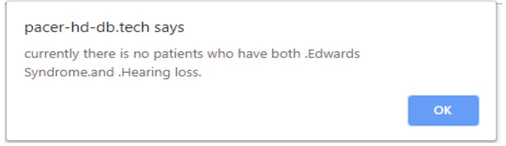
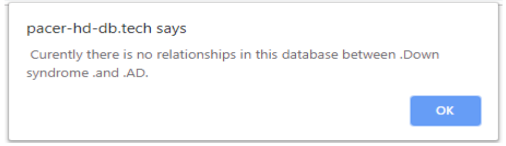
Please Select to Display Information about Patients:

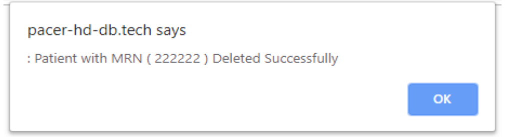
Figure 8.10: Interactive menus example in the display cohort page: three-option menu display

Customised alert messages

The system provides an interactive environment for the user as any action a user performs is considered in a case of success or failure. Accordingly, customised messages, as presented in Table 8.1, are designed to extract the message from the user’s query. The response is relevant to the user’s action and provides a satisfying answer. These types of messages appear in multiple cases in the search queries section, Query Builder pages. We implement a JavaScript function within the PHP to manage the alert message.

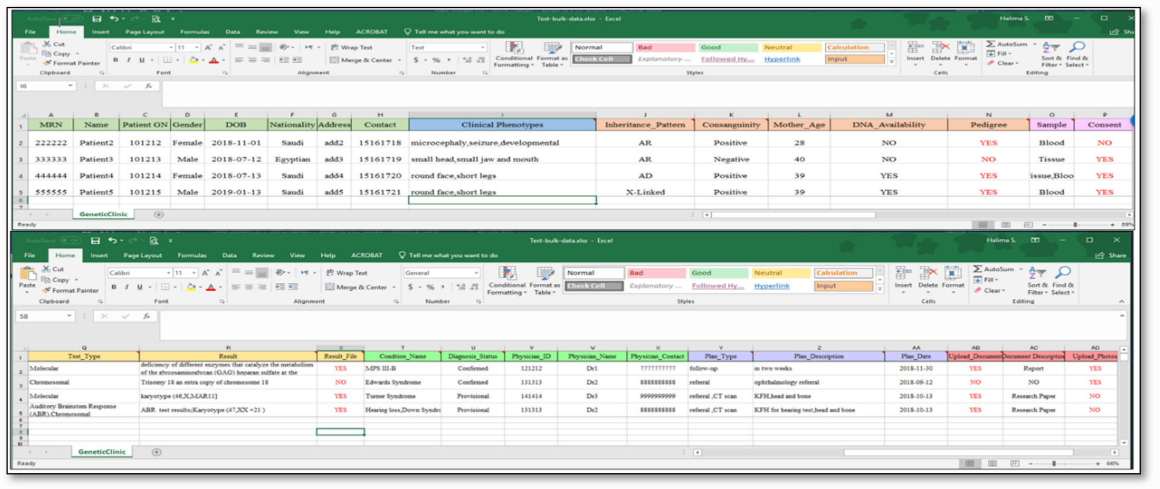
Table 8.1: List of alert messages

Explanation of the message	Alert message
When the page requests a patient MRN and to be checked, and the user tries to proceed without this step.	 pacer-hd-db.tech says Please Provide a Patient Medical Record Number (MRN).
When the user clicks on a Submit, **Display**, **Update**, **Remove**, etc. button with no selection made on the page.	 pacer-hd-db.tech says Please Select option(s)to display.
When the user tries to **Update** a patient by adding an existing clinical phenotype or a condition.	 pacer-hd-db.tech says You are trying to add an existing phenotype
When the user proceeds with the **Display** button leaving a required field for the query.	 pacer-hd-db.tech says You did not select a condtion for the mother age
When a user tried to click the **Upload** button and did not choose any file or attempted to upload a file with a different format, not Excel.	 pacer-hd-db.tech says Please Upload an Excel File type (xlsx)
When a user clicks the **Update** button to modify patient information by adding a new patient sample without consent.	 pacer-hd-db.tech says Please check the required fields Consent should be uploaded
When a user clicks the **Update** button to modify patient information by adding a new sample with the same name exists in the patient record.	 pacer-hd-db.tech says The patient has the same sample please use a different name Example (numbering the sample)
When the user selects to answer a query about displaying patients with specified two conditions, and there is no output, a customised message will appear according to the selected conditions.	 pacer-hd-db.tech says currently there is no patients who have both .Edwards Syndrome.and .Hearing loss.
When the user tries to answer a query about the relationships between a specified condition and inheritance patterns, but there is no output, a customised message will appear according to the selected condition and inheritance pattern.	 pacer-hd-db.tech says Curently there is no relationships in this database between .Down syndrome .and .AD.

A confirmation message when the deletion process is completed successfully.	 pacer-hd-db.tech says : Patient with MRN (222222) Deleted Successfully OK

Import functionality

The G3DMS offers another method of data entry in addition to the patient’s registration forms, which is to upload bulk data to migrate old data or data collected offline for research. Figure 8.11 shows the Excel file template, which is used to upload bulk data all at once. The template sheet provides the fields required by the database, which need to be filled with patient data in each row.



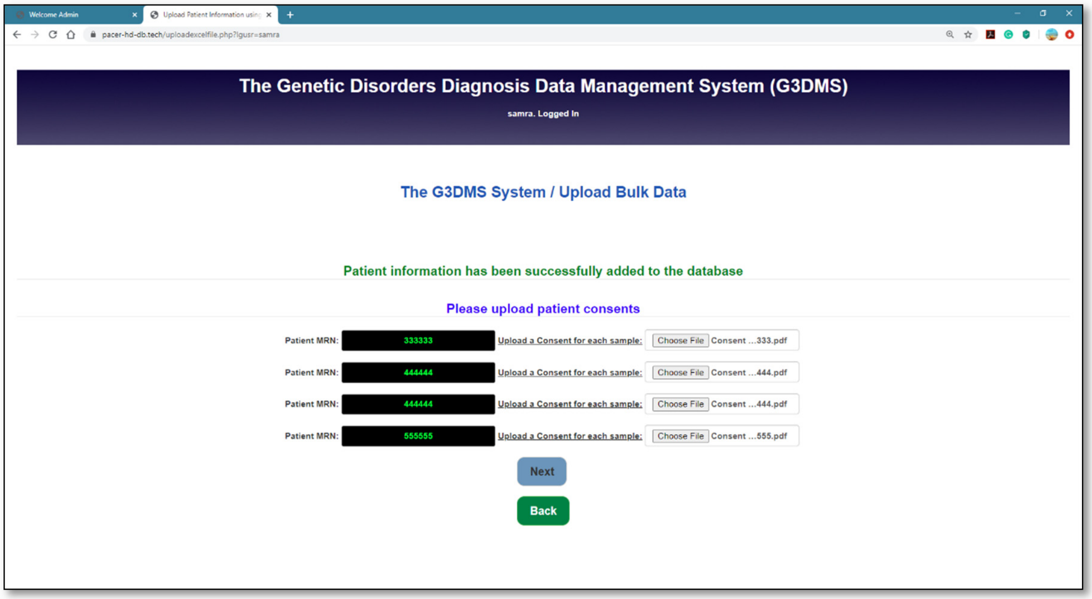
MRN	Name	Patient GN	Gender	DOB	Nationality	Address	Contact	Clinical Phenotypes	Inheritance_Pattern	Consanguinity	Mother_Age	DNA_Availability	Pedigree	Sample	Consent
222222	Patient2	101212	Female	2018-11-01	Saudi	add2	15161718	microcephaly,seizure,developmental	AR	Positive	28	NO	YES	Blood	NO
333333	Patient3	101213	Male	2018-07-12	Egyptian	add3	15161719	small head,small jaw and mouth	AR	Negative	40	NO	NO	Tissue	YES
444444	Patient4	101214	Female	2018-07-13	Saudi	add4	15161720	round face,short legs	AD	Positive	39	YES	YES	issue,Bloo	YES
555555	Patient5	101215	Male	2019-01-13	Saudi	add5	15161721	round face,short legs	X-Linked	Positive	39	YES	YES	Blood	YES

Test_Type	Result	Result_File	Condtion_Name	Diagnosis_Status	Physician_ID	Physician_Name	Physician_Contact	Plan_Type	Plan_Description	Plan_Date	Upload_Document	Document Descriptio	Upload_Photos
Molecular	deficiency of different enzymes that catalyze the metabolism of the glycosaminoglycan (GAG) heparan sulfate at the	YES	MPS III-B	Confirmed	121212	Dr1	7777777777	follow-up	in two weeks	2018-11-30	YES	Report	YES
Chromosomal	Trisomy 18 an extra copy of chromosome 18	NO	Edwards Syndrome	Confirmed	131313	Dr2	8888888888	referral	ophthalmology referral	2018-09-12	NO	NO	YES
Molecular	karyotype (46,X,MAR11]	YES	Turner Syndrome	Provisional	141414	Dr3	9999999999	referral ,CT scan	KFH,head and bone	2018-10-13	YES	Research Paper	NO
Auditory Brainstem Response (ABR),Chromosomal	ABR test results;Karyotype (47,XX +21)	YES	Hearing loss,Down Syndrc	Provisional	131313	Dr2	8888888888	referral ,CT scan	KFH for hearing test,head and bone	2018-10-13	YES	Research Paper	NO

Figure 8.11: Excel template to upload bulk data to the G3DMS

Mapping method

The upload button allows the loading of the Excel template and triggers the mapping process using PHPExcel classes as follows: require_once ‘Classes/PHPExcel/IOFactory.php’> The getActiveSheet() function is used to read the active worksheet and getHighestRow() and getHighestColumn() to get the content from each row and column and into an array of objects then allocate each value into their specific table location in the database. First, all the contents of the MRN column are stored in an array, then the array is looped, and each MRN is checked to determine if it exists in the database. If it does not, then the patient record is inserted in each field in its associated table, or else this patient is skipped. Within a patient “row” record, it is necessary to verify the presence of common field values such as conditions, clinical phenotypes and tests, in their separate table before inclusion, and then they are linked to patient tables. However, in the case of a multivalued column which should be separated by a comma, for example, a patient with two conditions, the mapping method recognises the separator and inserts each condition individually in the condition table after checking for its existence. If it exists, then the condition ID is retrieved, and it is referenced in the patient DIAGNOSIS table, or else the condition in the CONDITION table is inserted, and the ID is retrieved and linked to the DIAGNOSIS table. We follow a rule-based mapping technique for some columns, which are an image or file type, such as the patient (consent, results, pedigree chart, photo and documents). YES should fill each field if the patient's document is available or NO when there is no document for this field. All fields with text, date or a number value are inserted automatically, but for fields that require uploaded files and images, the system displays the upload options in each column for all patients on an individual page if the value is YES. Figure 8.12 shows the page for uploading the patient's consent filled with the YES value only, but, in the case of NO, the system discards this field and fills it with NULL. More illustration and detailed screenshots of how the mapping process is performed are provided in the user guide in Appendix 8.3.


Welcome Admin
Upload Patient Information using
pacer-hd-db.tech/uploadexcelfile.php?lgusr=samra
Next
Back

The Genetic Disorders Diagnosis Data Management System (G3DMS)

samra. Logged In

The G3DMS System / Upload Bulk Data

Patient information has been successfully added to the database

Please upload patient consents

Patient MRN: 333333 Upload a Consent for each sample: Choose File Consent ...333.pdf

Patient MRN: 444444 Upload a Consent for each sample: Choose File Consent ...444.pdf

Patient MRN: 444444 Upload a Consent for each sample: Choose File Consent ...444.pdf

Patient MRN: 555555 Upload a Consent for each sample: Choose File Consent ...555.pdf

Figure 8.12: Upload bulk data: patients' consents for YES value fields

Export functionalities

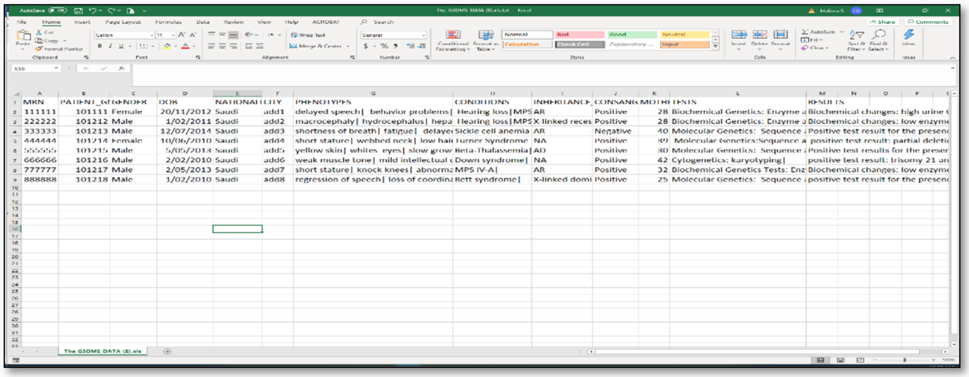
The G3DMS provides a customised mapping method for exporting the content of the database to an Excel file to be used for further research in the user's local organisation. The database content is also downloaded in a JSON file format, which can be integrated with other systems and Saudi open datasets. Only patient data which is useful for further research is exported. Other data, such as a patient's treatment plan and the doctor's information, are not required.

As the intention of the JSON file is to contribute to the research community outside the organisation, privacy rules are applied, such as removing a patient's identifiable information that is linked to the patient's identity.

Mapping method for exporting to Excel

For the export method to Excel file, a header function in PHP is used to create a new Excel file as follows: header("Content-Disposition: attachment; filename=The_G3DMS_DATA.xls")>

Then, the content is streamed from each table in the database (using foreach to navigate through the entire table), to be placed under each specified column header. A bar "|" is used as a delimiter in case of multivalued attributes such as clinical phenotypes, samples and tests. Figure 8.13 shows sample training data exported to Excel format. The exported data are restricted to the organisation that owns the data due to the provision of a patient's identifiable information.



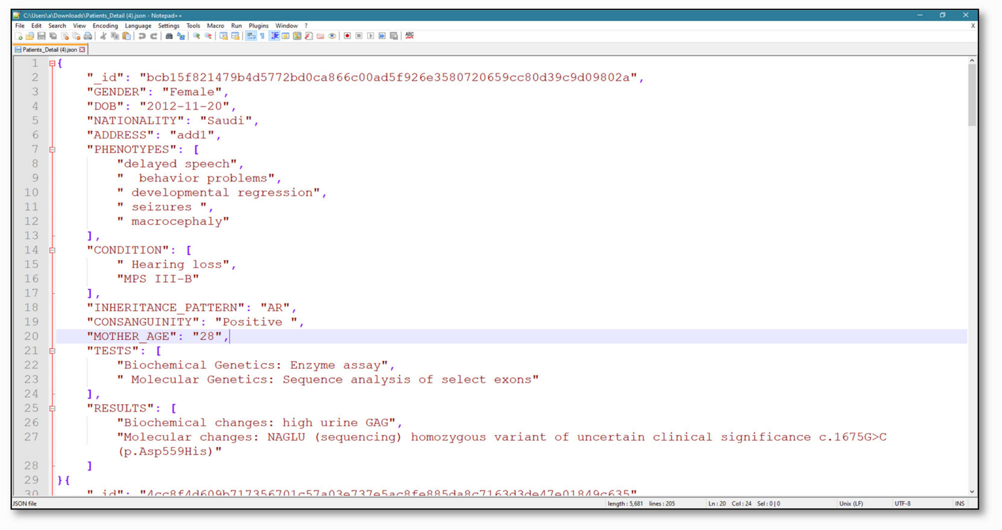
MRN	PATIENT_GI	GENDER	DOB	NATIONALI	CITY	PHENOTYPES	CONDITIONS	INHERITANCE_	CONSANG	MOTHI	TESTS	RESULTS
111111	101111	Female	20/11/2012	Saudi	add1	delayed speech| behavior problems|	Hearing loss|MPS	AR	Positive	28	Biochemical Genetics: Enzyme a	Biochemical changes: high urine
222222	101212	Male	1/02/2011	Saudi	add2	macrocephaly| hydrocephalus| hepa	Hearing loss|MPS	X-linked reces	Positive	28	Biochemical Genetics: Enzyme a	Biochemical changes: low enzyme
333333	101213	Male	12/07/2014	Saudi	add3	shortness of breath| fatigue| delaye	Sickle cell anemia	AR	Negative	40	Molecular Genetics: Sequence a	Positive test result for the presenc
444444	101214	Female	10/06/2010	Saudi	add4	short stature| webbed neck| low hai	Turner Syndrome	NA	Positive	39	Molecular Genetics:Sequence a	positive test result: partial deletic
555555	101215	Male	5/05/2013	Saudi	add5	yellow skin| whites eyes| slow grow	Beta-Thalassemia|	AD	Positive	30	Molecular Genetics: Sequence a	Positive test results for the preser
666666	101216	Male	2/02/2010	Saudi	add6	weak muscle tone| mild intellectual c	Down syndrome|	NA	Positive	42	Cytogenetics: karyotyping|	positive test result: trisomy 21 an
777777	101217	Male	2/05/2013	Saudi	add7	short stature| knock knees| abnorma	MPS IV-A|	AR	Positive	32	Biochemical Genetics Tests: Enz	Biochemical changes: low enzyme
888888	101218	Male	1/02/2010	Saudi	add8	regression of speech| loss of coordina	Rett syndrome|	X-linked domi	Positive	25	Molecular Genetics: Sequence a	positive test result for the presenc

Figure 8.13: Export data from the G3DMS into Excel file

Mapping method for exporting to JSON

A dedicated method is developed for read-only confirmed diagnosis cases and applies a de-identification method to remove the identifiable primary key, the MRN, to a new hashed key. Therefore, the data extracted can be used as valid datasets in the genetic conditions of the Saudi population without any concern about data privacy. First, a JSON file is created using the PHP header() function: header("Content-Disposition: attachment; filename=Patients_Detail.json");. Then, the extracted dataset is stored in an array of objects to be converted to a JSON representation using the json_encode() function. Figure 8.14 shows extracted data in the JSON file, the echo statement used to display the resulting encoded data into the created JSON file as follows:

```
echo json_encode($json_array,JSON_PRETTY_PRINT);.
```



```
{
    "_id": "bcb15f821479b4d5772bd0ca866c00ad5f926e3580720659cc80d39c9d09802a",
    "GENDER": "Female",
    "DOB": "2012-11-20",
    "NATIONALITY": "Saudi",
    "ADDRESS": "add1",
    "PHENOTYPES": [
        "delayed speech",
        "  behavior problems",
        " developmental regression",
        " seizures ",
        " macrocephaly"
    ],
    "CONDITION": [
        " Hearing loss",
        "MPS III-B"
    ],
    "INHERITANCE_PATTERN": "AR",
    "CONSANGUINITY": "Positive ",
    "MOTHER_AGE": "28",
    "TESTS": [
        "Biochemical Genetics: Enzyme assay",
        " Molecular Genetics: Sequence analysis of select exons"
    ],
    "RESULTS": [
        "Biochemical changes: high urine GAG",
        "Molecular changes: NAGLU (sequencing) homozygous variant of uncertain clinical significance c.1675G>C
        (p.Asp559His)"
    ]
}{
    "_id": "4cc8f4d609b717356701c57a03e737e5ac8fe885da8c7163d3de47e01849c635"
```

Figure 8.14: Export the dataset from the DESIGNED system into a JSON file

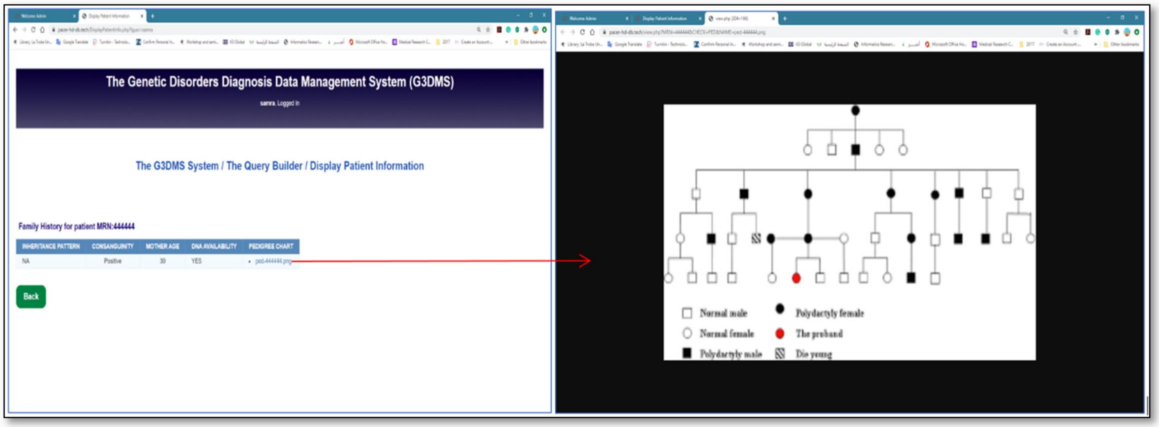
File and image management

One of the critical requirements of the G3DMS is the ability to accommodate patient-related documents, such as consents, photos, pedigree charts, results, reports and research papers. Therefore, the focus was to deliver a secure method for uploading, storing and retrieving this type of data. The decision was to store images and files in the database as a binary object BLOB instead of in a file system storage. As these data files can be used in research queries, they need to be stored in the same database with the related patient's data to maintain consistency and data integrity. First, after submitting the uploaded file, the PHP $_POST method passes the content to the PHP file to get the file name, type and content to be stored in three separate variables, for example, storing the pedigree chart image for a patient:

```
$pedname = $_FILES['ped']['name'];
$pedtype = $_FILES['ped']['type'];
$ped_Data = file_get_contents($_FILES['ped']['tmp_name']);
$pedContent= addslashes($ped_Data);
```

Next, insert the file into the FAMILY_HISTORY table. To display an image or a PDF file, as shown in Figure 8.15, the PHP header() function is implemented as follows:

```
$sql = "SELECT * FROM `family_history` WHERE `MRN`=$id and `PEDIGREE_NAME` like '%$name%'";
 $result = mysqli_query($conn, $sql) or die ("Bad Query:$sql"); while($row =mysqli_fetch_assoc($result))
{ header('Content-Type:'.$row['PEDIGREE_TYPE']); echo $row['PEDIGREE_CONTENT'];}
```



The Genetic Disorders Diagnosis Data Management System (G3DMS)

samra. Logged In

The G3DMS System / The Query Builder / Display Patient Information

Family History for patient MRN:444444

INHERITANCE PATTERN	CONSANGUINITY	MOTHER AGE	DNA AVAILABILITY	PEDIGREE CHART
NA	Positive	39	YES	• ped-444444.png

Figure 8.15: Display Patient Information: the result page and view pedigree

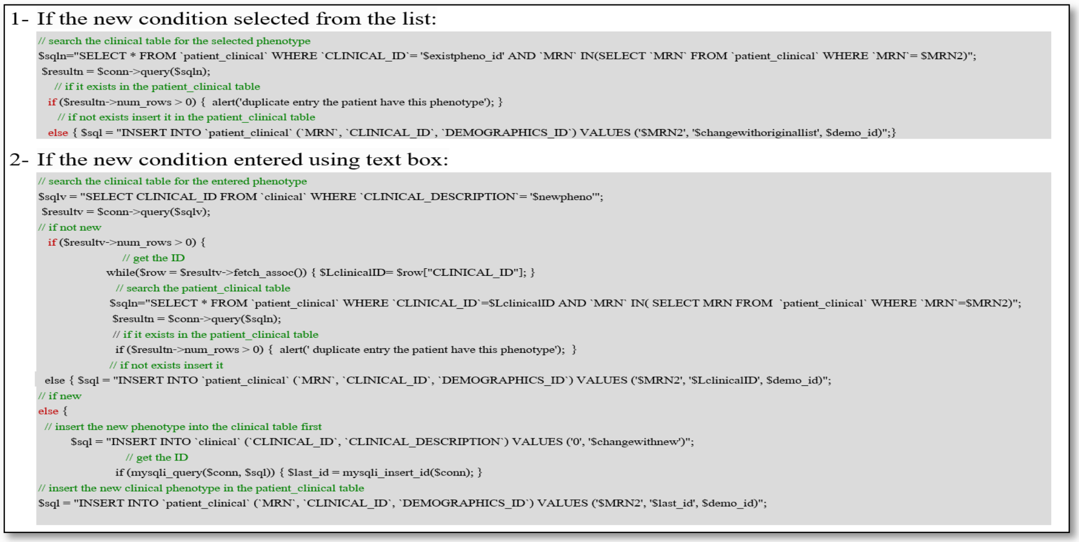
Simple query

The G3DMS provides query services in multiple forms. Simple queries are used during patient registration for checking new entries such as patient MRN, phenotypes and conditions before performing insertion into the database to prevent data duplication. A simple search option is also used to search for a specific patient using the MRN and displays the patient information (demographic information, clinical phenotypes, diagnosis, family history, samples and consents, tests and results, plans, photos and documents). This can be achieved using a simple select statement, and then the results are displayed using an HTML table as illustrated in the display results section. Similarly, a simple search operation is performed to delete a patient for the given MRN, then the demographic information for this patient is retrieved to double-check their identity and confirm the delete action to remove all the patient's records from the database. The following is an example of a simple query for checking the MRN to delete, update or display the patient information:

```
$query = mysqli_query($conn, "SELECT * FROM patient WHERE MRN='".$PATIENT_MRN."'");
if(mysqli_num_rows($query3) > 0){  $Exist ="The medical record number exists in the database, please select
the required action";   }
  else{   $notExist= "The medical record number does not exist, please try again"; }
```

Complex query

Complex actions such as modifying patient information by removing existing or adding new information require a more advanced method to perform all the necessary internal operations before completing the task. The action of updating the patient clinical phenotypes by adding or removing them is performed in several steps. Figure 8.16 shows the script for updating the patient's information by adding a new clinical phenotype for a given patient's MRN, and the user can enter the new phenotype from the database list or use the text box to type in the phenotype.



1- If the new condition selected from the list:

```
// search the clinical table for the selected phenotype
$sqln="SELECT * FROM `patient_clinical` WHERE `CLINICAL_ID`= '$existpheno_id' AND `MRN` IN(SELECT `MRN` FROM `patient_clinical` WHERE `MRN`= $MRN2)";
 $resultn = $conn->query($sqln);
    // if it exists in the patient_clinical table
  if ($resultn->num_rows > 0) {  alert('duplicate entry the patient have this phenotype'); }
     // if not exists insert it in the patient_clinical table
  else { $sql = "INSERT INTO `patient_clinical` (`MRN`, `CLINICAL_ID`, `DEMOGRAPHICS_ID`) VALUES ('$MRN2', '$changewithoriginallist', $demo_id)";}
```

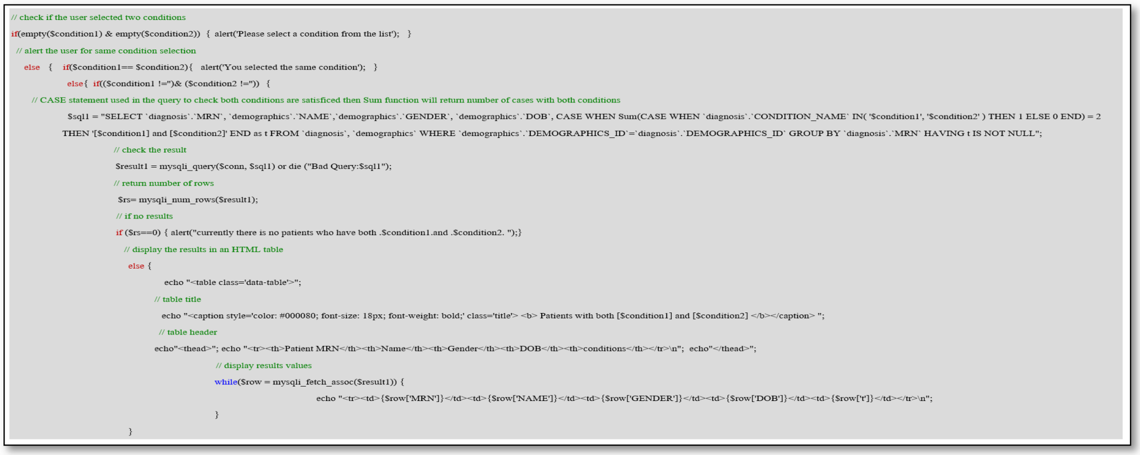
2- If the new condition entered using text box:

```
// search the clinical table for the entered phenotype
$sqlv = "SELECT CLINICAL_ID FROM `clinical` WHERE `CLINICAL_DESCRIPTION`= '$newpheno'";
 $resultv = $conn->query($sqlv);
// if not new
  if ($resultv->num_rows > 0) {
                        // get the ID
                 while($row = $resultv->fetch_assoc()) { $LclinicalID= $row["CLINICAL_ID"]; }
                       // search the patient_clinical table
                  $sqln="SELECT * FROM `patient_clinical` WHERE `CLINICAL_ID`=$LclinicalID AND `MRN` IN( SELECT MRN FROM  `patient_clinical` WHERE `MRN`=$MRN2)";
                   $resultn = $conn->query($sqln);
                   // if it exists in the patient_clinical table
                   if ($resultn->num_rows > 0) {  alert(' duplicate entry the patient have this phenotype');  }
                  // if not exists insert it
  else { $sql = "INSERT INTO `patient_clinical` (`MRN`, `CLINICAL_ID`, `DEMOGRAPHICS_ID`) VALUES ('$MRN2', '$LclinicalID', $demo_id)";
// if new
else {
 // insert the new phenotype into the clinical table first
       $sql = "INSERT INTO `clinical` (`CLINICAL_ID`, `CLINICAL_DESCRIPTION`) VALUES ('0', '$changewithnew')";
                      // get the ID
                   if (mysqli_query($conn, $sql)) { $last_id = mysqli_insert_id($conn); }
// insert the new clinical phenotype in the patient_clinical table
$sql = "INSERT INTO `patient_clinical` (`MRN`, `CLINICAL_ID`, `DEMOGRAPHICS_ID`) VALUES ('$MRN2', '$last_id', $demo_id)";
```

Figure 8.16: Script for updating patient information by adding a new clinical phenotype

I. Advanced query

Specific queries are implemented to answer some research questions displayed in general form and can be controlled by the user selection to be executed in more complex queries. Figure 8.17 illustrates a common research question where the user can identify two related or suspected conditions in one way or another and show whether a patient in the database suffers from both conditions. Two CASE statements are implemented. First, there is an inner CASE to check that both conditions are in the patient diagnosis table. If it is satisfied, it returns 1, and the Sum function is used to count the number of elements inside the CASE statement. If the results are equal to 2, then both conditions are combined with “and” in another variable “t” and grouped by patient MRN and having “t” is not null, if all conditions are satisfied the second outer CASE will return the results.



```
// check if the user selected two conditions
if(empty($condition1) & empty($condition2))  {  alert('Please select a condition from the list');   }
  // alert the user for same condition selection
    else   {    if($condition1== $condition2){   alert('You selected the same condition');   }
                 else{  if(($condition1 !='')& ($condition2 !=''))  {
     // CASE statement used in the query to check both conditions are satisficed then Sum function will return number of cases with both conditions
                $sql1 = "SELECT `diagnosis`.`MRN`, `demographics`.`NAME`,`demographics`.`GENDER`, `demographics`.`DOB`, CASE WHEN Sum(CASE WHEN `diagnosis`.`CONDITION_NAME` IN( '$condition1', '$condition2' ) THEN 1 ELSE 0 END) = 2
               THEN '[$condition1] and [$condition2]' END as t FROM `diagnosis`, `demographics` WHERE `demographics`.`DEMOGRAPHICS_ID`=`diagnosis`.`DEMOGRAPHICS_ID` GROUP BY `diagnosis`.`MRN` HAVING t IS NOT NULL";
                          // check the result
                          $result1 = mysqli_query($conn, $sql1) or die ("Bad Query:$sql1");
                          // return number of rows
                           $rs= mysqli_num_rows($result1);
                           // if no results
                          if ($rs==0) { alert("currently there is no patients who have both .$condition1.and .$condition2. ");}
                            // display the results in an HTML table
                             else {
                                       echo "<table class='data-table'>";
                                   // table title
                                      echo "<caption style='color: #000080; font-size: 18px; font-weight: bold;' class='title'> <b> Patients with both [$condition1] and [$condition2] </b></caption> ";
                                    // table header
                                   echo"<thead>"; echo "<tr><th>Patient MRN</th><th>Name</th><th>Gender</th><th>DOB</th><th>conditions</th></tr>\n";  echo"</thead>";
                                                    // display results values
                                                    while($row = mysqli_fetch_assoc($result1)) {
                                                                          echo "<tr><td>{$row['MRN']}</td><td>{$row['NAME']}</td><td>{$row['GENDER']}</td><td>{$row['DOB']}</td><td>{$row['t']}</td></tr>\n";
                                                    }
                             }
```

Figure 8.17: Advanced query to display patients with two selective conditions

8.3 Documentation phase

The documentation phase collects the necessary documentation from the system design lifecycle stages and prepares a user guide to support the end-users' use of the system.

8.3.1 System documentation

System documentation includes documents that describe the actual system structure and all the material involved in the development of the system throughout the design lifecycle. The product of each stage is assigned in the systems development lifecycle to document part of the system design. Therefore, for requirements documentation, the reference is the strategy and analysis phases. The design architecture documents and the source code can be referenced in the design and build phases. The verification test documents are prepared during the transformation phase.

8.3.2 User documentation

The user-centred document guide for the end-users is delivered through user-friendly instructional content in Appendix 8.3. The aim is to display in detail all the instructions for using the system, how to deal with it, what to expect, how it will react to errors, and how errors are handled using relevant, informative messages. The documentation is organised according to user tasks and system usage scenarios. It includes screenshots of inputs and outputs with a complete description of each part of the system, including errors and messages to address these errors. User guide documentation in a PDF file is provided as a downloadable attachment via a link in the G3DMS home pages.

8.4 Transition phase

This phase provides a smooth transition for the physical database from the development and testing environment to production and implementation. The complete system, including the backend database and the system application interface, is subject to validation testing to ensure the system meets user requirements. This phase encompasses data loading, conversion and end-user training. The end-user documentation is used as training material for end-users to refer to if they require further instruction to complete their tasks and answer their research enquiries.

8.4.1 System testing

The system supports user interaction with the database through forms with various fields. Therefore, the goal is to test the accuracy of the frontend (database application interface) against the backend database as well as to verify the accuracy of the database in how data are stored and accessed in the database. A validation test was performed to ensure that the application works correctly according to the requirements and rules specified in the strategy and analysis phases. Sample data are used to test the application interface against the database. The sample data hold real data characteristics except for false MRN and demographic data. Under our supervision, a member of the PACER-HD conducted a system validation test by trying several case scenarios using a testing checklist. Database testing was performed using phpMyAdmin as a testing environment. The testing outcomes were very satisfactory.

Interface testing checklist

☑ Validate mandatory fields

☑ Validation error messages

☑ General confirmation messages

☑ Validate forms entry to the database

☑ Importing bulk data upload functionality

☑ Display patient information

☑ Display cohort

☑ Update patient information (Remove, Add)

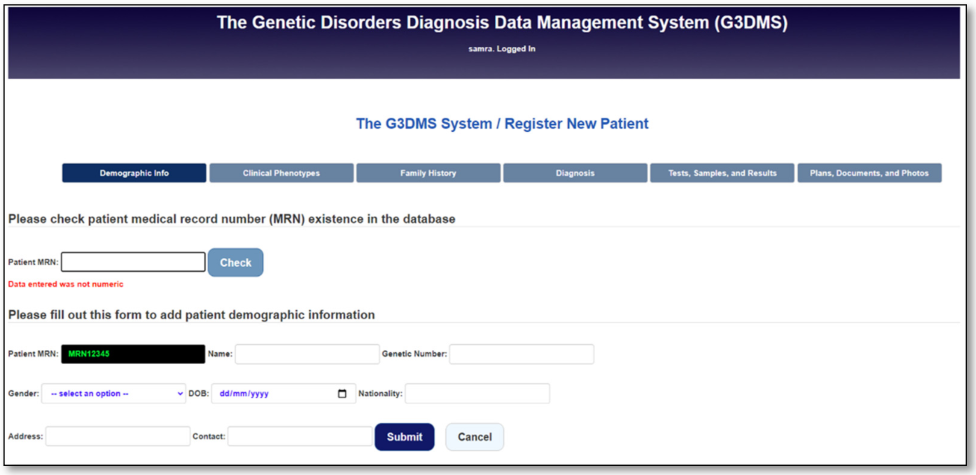
☑ Delete patient record

☑ Research queries functionality

☑ Export the database

Examples

Field entry validation messages are displayed for each field to prevent the insertion of blank and non-text variables into the field. The field MRN was tested with a non-numeric entry, and the message “Data entered was not numeric” was displayed as a validation action, as shown in Figure 8.18. Figure 8.19 shows an example of a confirmation message posted when the user uploaded bulk data to the database.



The Genetic Disorders Diagnosis Data Management System (G3DMS)

samra. Logged In

The G3DMS System / Register New Patient

Please check patient medical record number (MRN) existence in the database

Patient MRN:

Data entered was not numeric

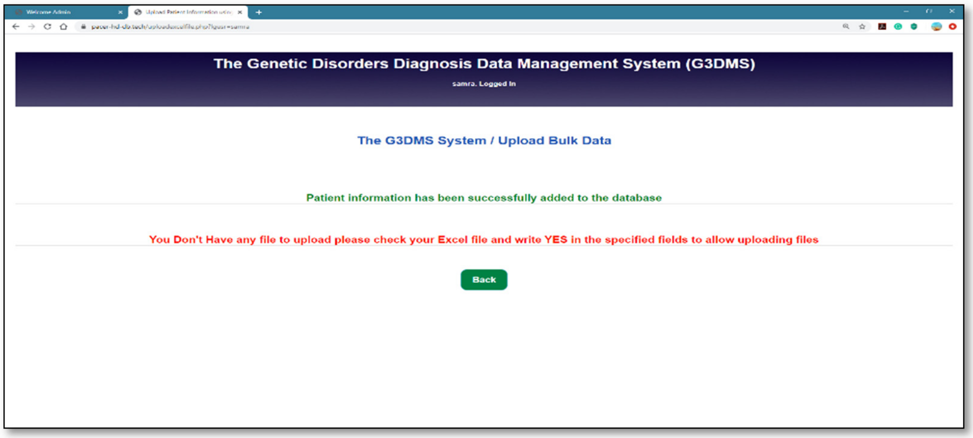
Please fill out this form to add patient demographic information

Patient MRN: MRN12345 Name: Genetic Number:

Gender: -- select an option -- DOB: dd/mm/yyyy Nationality:

Address: Contact:

Figure 8.18: Field entry validation message



The Genetic Disorders Diagnosis Data Management System (G3DMS)

samra. Logged In

The G3DMS System / Upload Bulk Data

Patient information has been successfully added to the database

You Don't Have any file to upload please check your Excel file and write YES in the specified fields to allow uploading files

Figure 8.19: Confirmation messages

8.4.2 Database testing

We performed several test scenarios to test the database for accuracy in data storage and retrieval. The testing is conducted at the development and testing environment with phpMyAdmin as an administration tool to run the queries. The following checklist is used to validate data entry, storage and access in the database.

Database testing checklist

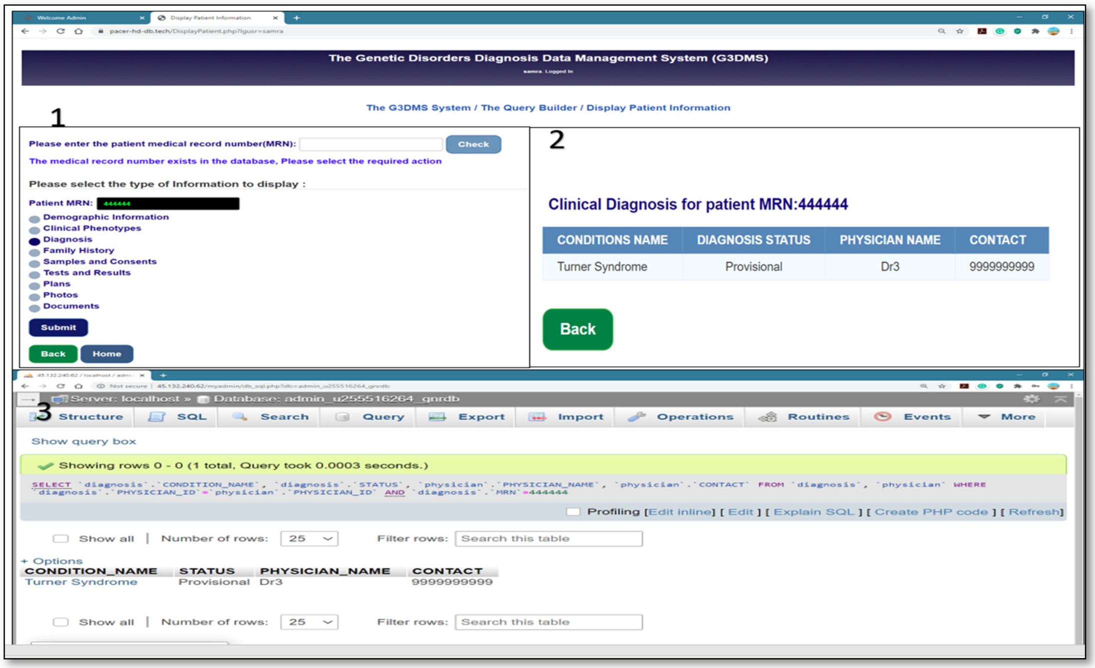
- ☑ Data allocated correctly to the right fields in the database upon successful page submission.
- ☑ Preserve data integrity when inserting, deleting or updating information in the database.

- ☑ Data entry in the database is sequential. First, the patient record with the demographic information should be created before filling out any other information.
- ☑ Mandated data entry fields saved in primary key positions in the database.
- ☑ Check uploaded images and files stored properly in the corresponding fields (name, type and content) database.
- ☑ Check the functionality of the uploaded file of the supported type (PDF) and the image extensions (PNG, BMP, JPEG.)
- ☑ Check the validity of the link to display files and images.
- ☑ Check the quality of the retrieved image from the database.
- ☑ Check query results, specified columns and correct values are displayed.

Examples

The database was tested for a query using the selection of display patient diagnosis information in the G3DMS interface and the following SQL script for the query in phpMyAdmin, as shown in Figure 8.20:

```
SELECT`diagnosis`.`CONDITION_NAME`,`diagnosis`.`STATUS`,`physician`.`PHYSICIAN_NAME`,
`physician`.`CONTACT`            FROM            `diagnosis`,            `physician`            WHERE
`diagnosis`.`PHYSICIAN_ID`=`physician`.`PHYSICIAN_ID` AND `diagnosis`.`MRN`=444444;
```


Welcome Admin
Display Patient Information
pacer-hd-db.tech/DisplayPatient.php?lgusr=samra
1
2
3

The Genetic Disorders Diagnosis Data Management System (G3DMS)

samra. Logged In

The G3DMS System / The Query Builder / Display Patient Information

Please enter the patient medical record number(MRN): Check

The medical record number exists in the database, Please select the required action

Please select the type of Information to display :

Patient MRN: 444444

○ Demographic Information
○ Clinical Phenotypes
● Diagnosis
○ Family History
○ Samples and Consents
○ Tests and Results
○ Plans
○ Photos
○ Documents

Submit

Back Home

Clinical Diagnosis for patient MRN:444444

CONDITIONS NAME	DIAGNOSIS STATUS	PHYSICIAN NAME	CONTACT
Turner Syndrome	Provisional	Dr3	9999999999

Back

45.132.240.62 / localhost / adm
Not secure | 45.132.240.62/myadmin/db_sql.php?db=admin_u255516264_gnrdb
Server: localhost » Database: admin_u255516264_gnrdb
Structure
SQL
Search
Query
Export
Import
Operations
Routines
Events
More

Show query box

✓ Showing rows 0 - 0 (1 total, Query took 0.0003 seconds.)

```
SELECT `diagnosis`.`CONDITION_NAME`, `diagnosis`.`STATUS`, `physician`.`PHYSICIAN_NAME`, `physician`.`CONTACT` FROM `diagnosis`, `physician` WHERE
`diagnosis`.`PHYSICIAN_ID`=`physician`.`PHYSICIAN_ID` AND `diagnosis`.`MRN`=444444
```

□ Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

□ Show all | Number of rows: 25 Filter rows: Search this table

+ Options

CONDITION_NAME	STATUS	PHYSICIAN_NAME	CONTACT
Turner Syndrome	Provisional	Dr3	9999999999

□ Show all | Number of rows: 25 Filter rows: Search this table

Figure 8.20: Testing the query (display patient diagnosis information) shows equivalent results

8.5 Production phase

Production is the last design stage, which is the application and deployment stage. The content of the system is moved from the test environment, the XAMPP platform, to a live server host using a commercial host for the domain, website and hosting servers. The website https://pacer-hd-db.tech/ is used to access the application interface of the G3DMS online. While the database can be accessed and managed through the phpMyAdmin interface, any changes and updates to the application pages can be done locally at the PACER-HD centre and then uploaded using FileZilla software. The users have access to the website using their login account as well as access to the database content through phpMyAdmin at their host server.

8.6 Qualitative evaluation

Good system design is important to ensure system functions work in accordance with their intended purpose. A well-implemented system also affects user satisfaction in relation to sustainable adoption [9]. A poor user and organisational contribution in relation to the design process and implementation may mean the potential of a health information system is not realised. A lack of clinical engagement is one of the important reasons behind physician resistance which is one of the most crucial barriers to health information system adoption. Therefore, two evaluation methods were adopted to consolidate the design and ensure the successful implementation of the G3DMS. First, the design and implementation of the system follow the multilevel service design approach, which fosters user adoption at the implementation stage [148]. Also, an informatics evaluation framework provides a heuristic for matching the stage of system development according to the system design lifecycle and the level of evaluation [149]. Both methods of evaluation increase user experience with the system, and constant feedback is given in an iterative design lifecycle.

8.6.1 Multilevel service evaluation method

Inspired by the successful development and implementation based on a service design approach of a widely used Portuguese electronic health record, a qualitative study performed after implementation showed that the electronic health record was considered useful and easy to use, and these results are backed by the widespread use of the system [147]. Therefore, this thesis uses the multilevel service design approach to evaluate the systems development lifecycle of the design and implementation of the G3DMS, where all the principles of the approach are delivered through the Barker method stages of the systems development lifecycle. The aim is to evaluate the extent of user involvement in the design and gain a holistic view of the complete design process from service concept level to the multi-interface service system level and to each service encounter. The evaluation is based on the four process steps of the multilevel service design: studying the system user experience, designing the service concept, designing the service system, and designing the service encounter. Table 8.2 shows the steps of the multilevel service design approach with the evaluation of the corresponding action in the G3DMS systems development lifecycle. The evaluation guaranteed the involvement of the system user throughout the design and implementation process, which will increase user familiarity with and acceptance of the system.

Table 8.2: Evaluation of the G3DMS using the multilevel service design approach

MSD Steps	G3DMS application of user experience	Evaluation of service offering
Step 1: Study the system user's experience	Case study: PACER-HD Focus group, brainstorming method for discussing system requirements (process and data)	Understanding process and data requirements provide the basis for designing a service offering
Step 2: Design the service concept	Strategy phase of the Barker method for SDLC Display (the basic entity-relationship diagram and the data flow diagram level zero)	Analysis of the existing diagnosis process flow and the data elements provided by the users of the system resulted in the basic entity-relationship diagram and the data flow diagram level zero.
Step 3: Design the service system	The analysis phase of the system design lifecycle Display (the global schema, and the data flow diagram level 1)	Use of modelling techniques such as using the entity-relationship diagram for the global schema and data flow diagram for the process flow provided a visual interaction of the design, so the users can evaluate the models based on their experience.
Step 4: Design the service encounter	Design, build, documentation, transition and production phases of SDLC Display (the logical schema, site map, the wireframe diagram, user manual)	Based on the user experience on the diagnosis process workflow and requirements for research, the blueprint presented, the application interface in a web pages style using the site map, wireframe diagram, and user manual for illustrating each page content and usage.

8.6.2 Informatics evaluation framework

Table 8.3 shows the five stages of the system development lifecycle and equivalent evaluation level according to the Stead et al. framework, as presented by Kaufman et al. [149]. The results of one level of evaluation can be used at a later stage of development or can indicate the need to move to the next level of evaluation. Adopting this model allows the evaluation process to reflect knowledge in the development process.

Table 8.3: An informatics evaluation framework for G3DMS, based on Kaufman et al. [149], adopted from Stead et al. [208]

Stage of System Development	Level of System Evaluation	G3DMS Evaluation Method
Stage 1: Specification and Needs Requirement	Evaluate specifications	A brainstorming method is carried out with a focus group from PACER-HD, including the head of the organisation, four physicians and the researcher. The purpose is to evaluate and define the problem and acquire the requirements (process requirements and data requirements).
Stage 2: Component Development (the database and the user interface)	Evaluate in the lab	Unit testing (database and the user interface) performed in the development environment (XAMPP platform) using MySQL module for database testing through phpMyAdmin and the application interface in the Apache server module through the localhost at port: 8080. The aim of these tests is to evaluate the programming code functionality of the user interface and query performance in the database.
Stage 3: Combine Components		The whole system interaction (user-interface and the database) is tested in the same environment. The goal of these tests is to evaluate the interaction and response between the frontend user interface and the backend database.
Stage 4: Integration of Components into System	Evaluate in the field	The completed version of the system was tested in the local environment (researcher's laptop) at PACER-HD by two physicians under our observation, and feedback notes were taken during the trial session. The purpose of this evaluation is to capture feedback and observe the user interaction with the system in term of difficulties and ease of use.
	Evaluate validity	Following the system deployment on the internet, a link to the system is provided to the designated physician from PACER-HD with a username and password. A validation test checklist is also emailed to help with the assessment as well as the user manual which can explain the functionality of the interface pages and all the messages that may appear during user interaction with the system. The aim of this evaluation is to allow system users to test the usability of the system without our influence.
Stage 5: Routine Use of a System	Evaluate efficacy	The system is at its early stage as the PACER-HD management planning for transforming their legacy data to G3DMS. After using the system for three months, the FITT framework will be used to evaluate the user, system and process interaction using the appropriate method for data collection (either quantitative, interviews or focus group) to evaluate the system. The purpose of this evaluation is to determine the effectiveness of the system in the care process and for research studies.

8.6.3 Success factors

- Physicians are involved in the system planning and design lifecycle (iterative, user provides comments on the trial version, which are used to alter the design model).
- The system users tested G3DMS validity according to their requirements, and it was found to support their workflow.
- The system is cost-effective. It is a free system; that is, the department does not have to pay for the system as we offered free training and assisted with legacy data transformation.
- It is subject to future development and support as the centre is part of the university hospital at King Abdulaziz University, where the researcher teaches.

8.6.4 Recommendation for successful implementation and sustainability

Recommendations include to:

- Have a managerial role to assign users to upload the legacy data into the system to save time.
- Assign administrative staff from the department to manage the service provider account payments.
- Ensure IT staff supervise the database and website issues.
- Hold regular meetings with IT, staff and physicians to discuss issues related to system usage.

8.6.5 Action plan evaluation

Table 8.4 shows the process of evaluating the action plan in terms of goal, team members, challenges of the current system in PACER-HD, diagnosis workflow of G3DMS, and the design and development process of G3DMS.

Table 8.4: Action plan evaluation

<table>
<tr><th colspan="4">Goal</th></tr>
<tr><td colspan="4">Design and Implement a Data Management System for the Diagnosis of Genetic Disorders</td></tr>
<tr><th>Description</th><th colspan="2">Team Members</th><th>Role</th></tr>
<tr><td rowspan="7">Planning team</td><td colspan="2">Mrs Halima Samra, PhD candidate, La Trobe University, Melbourne, Australia</td><td>System developer</td></tr>
<tr><td colspan="2">Prof. Ben Soh, Associate Professor, La Trobe University, Melbourne, Australia</td><td>Supervisor</td></tr>
<tr><td colspan="2">Dr Alice Li, Senior Lecturer, La Trobe University, Melbourne, Australia</td><td>Co-Supervisor</td></tr>
<tr><td colspan="2">Chair of Dept of Genetic Medicine, Faculty of Medicine and Director of PACER-HD—Jeddah-KSA</td><td>Higher Management</td></tr>
<tr><td colspan="2">A researcher at PACER-HD</td><td>Superuser</td></tr>
<tr><td colspan="2">A research assistant and lab coordinator at PACER-HD</td><td>System user</td></tr>
<tr><td colspan="2">Three physicians who are researchers at PACER-HD</td><td>System user</td></tr>
<tr><th>Description</th><th>Diagnosis Workflow</th><th>Outcomes</th><th>For Research Study Purpose</th></tr>
<tr><td>The current system challenges in PACER-HD</td><td>All the process was done manually, and patient data were stored in a paper-based format (test results, reports, photos and research papers).</td><td>An unwieldy way of reading all paper-based documents and drawing conclusions for the purpose of patient diagnosis. It is almost impossible to compare data of multiple patients to confirm a diagnosis.</td><td>All data collected manually then entered into Excel files for analysis. However, it is difficult to answer complex research questions.</td></tr>
<tr><td>G3DMS</td><td>Support standardised data collection using electronic forms for structured data collection, storage and processing, as well as allowing PDF files and all types of images to be uploaded.</td><td>Easy retrieval of patient data and access to all data and files electronically with the possibility of reviewing them with the results of other patients.</td><td>Allow uploading old data from Excel. Provide query interface pages to support simple, advanced and complex research questions, in addition to export data from the database in Excel and JSON format.</td></tr>
<tr><th colspan="4">Action plan for the design and development of the G3DMS</th></tr>
<tr><th>Task Description</th><th>Member</th><th>Resources</th><th>Evaluation</th></tr>
<tr><td>Problem definition, requirements and system specification</td><td>System developer, higher management, super users and other users from research groups</td><td>Members meeting in a focus group discussion</td><td>The task accomplished its goal and was completed on time without delay</td></tr>
<tr><td>Prepare conceptual frameworks</td><td>System developer</td><td>Present the model to the superuser for confirmation of data element and diagnosis process flow</td><td>The task outcome was satisfactory with minor amendments on data elements to suit the patient-centred design approach</td></tr>
<tr><td>Prepare the logical frameworks</td><td>System developer</td><td>Present the model to the superuser for confirmation of data element and diagnosis process flow</td><td>The resulting blueprint was clearly understood by the user</td></tr>
<tr><td>Programming and coding for both database and user interface</td><td>System developer</td><td>Development environment in the developer personal computer</td><td>Iterative testing until developer satisfaction</td></tr>
<tr><td>Documentation</td><td>System developer</td><td>Step-by-step action scenarios with screenshots converted to a PDF file</td><td>This task resulted in a useful user guide</td></tr>
</table>

Implementation and system deployment	System developer	Hiring a virtual private server; superuser and other users for testing the system; sample data for testing	Satisfactory outcomes, no errors, appropriate messages, accurate results
Supervision and review	Supervisor and Co-supervisor	Each stage documentation and coding	Approval of submitted work for each phase

8.7 Results

8.7.1 G3DMS

A case study was undertaken to analyse a health information system in Saudi to understand its design problems via a brainstorming method with a focus group from the Princess Al-Jawhara Centre of Excellence in Research of Hereditary Disorders, King Abdulaziz University Hospital, Jeddah. The Barker system design method and a prototype were used to validate the proposed system via usability testing and a health informatics validation framework. The G3DMS comprises electronic data capture forms for data entry; a customised query builder to display and modify patient data as well as form research queries; a module that allows historical data to be uploaded in the form of bulk data using a template; export data options to Excel and JavaScript Object Notation format; and authorisation access for healthcare researchers and clinicians. The G3DMS was implemented in the Princess Al-Jawhara Centre of Excellence in Research of Hereditary Disorders at King Abdulaziz University Hospital.

8.7.2 Design features

During planning for the design of the data management system, several aspects were considered to reach the ultimate goal of an effective system that serves the purpose, achieving the satisfaction of the system users, and ensuring successful implementation and sustainability in use. First, adhere to the health informatics specification for the healthcare system to be able to support the workflow, assess with making decisions, and allow reuse of data collected by such systems. Second, use a flexible and comprehensive method for the design lifecycle of the system which covers all design stages in detail with the Barker method and adopt a service design approach to ensure system users' participation from system planning to validation and evaluation. Finally, monitor the design lifecycle stages using the informatics evaluation framework with iterative steps of feedback and reform.

8.7.2.1 Informatics-enabled framework

The main objective of the G3DMS is aiding clinicians at the point of care by optimising the diagnosis workflow and allowing informed decision to confirm conditions based on the collective data through the system and provide a query interface to allow researchers to customise their research questions based on the data available in the system.

Support the diagnosis workflow: The system is designed to capture patient data within the process flow of the diagnosis, as illustrated in Figure 7.2. The application interface is designed to present each step with an equivalent web page (s) which contain an electronic form for standardised data collection and enforce required fields to ensure data completeness. Self-developing dynamic lists are automatically updated during data entry which saves physician time and allows them to focus more on patient examination. The web pages flow dynamically according to the new patient diagnosis workflow; each finished page leads to the next page until the end of the diagnosis process. However, for existing patients, any information can be modified using patient MRN and by selecting from the category list in the update patient page, either to add, change or remove patient information.

Support physician in making diagnosis decision: The system is designed to allow uploading documents (test results, reports and published papers) and patient photos which can be used to assist with diagnosing genetic conditions. The diagnosis workflow has two different paths according to the diagnosis status (provisional or confirmed), as shown in Figure 7.2. In provisional diagnosis, usually, the doctor requires more investigative genetic or non-genetic testing, then test results with other supportive documents such as reports, publications and patient photos can be retrieved from the system to assist with confirming the diagnosis and help the physician to make an informed decision.

Support researchers in performing research studies: The system is designed to support research studies in multiple forms. Electronic capture forms for data collection ensure data quality and completeness. Progressive lists automatically filled with frequent entered common phenotypes, conditions and test results prevent data duplication. Query builder interface pages were specifically designed to allow researchers to navigate the content of the system for the individual patient as well as cohorts using customised queries and questions and to present the results in a table format supported with some statistics such as grouping the results according to patients' gender, condition, phenotype and location address. The researchers can also obtain and share unidentified datasets from the system using customised methods to automatically remove the patient's identifiable information before exporting to Excel or JSON format.

8.7.2.2 Effective design methods

Applying a case study allowed us to focus on the requirements and put the design strategy alongside the potential users of the system in a focus group meeting to discuss the requirements from the system developer and users' views. Revising the possible design approaches, we then decided to adopt methods that allow user involvement in the design process. The chosen method should be easy to represent the ideal working environment, provide models with understandable terms easy to communicate to users and allow iterative steps to consider user feedback and response. For these reasons, the design method for the G3DMS design lifecycle included the Barker method to facilitate the design lifecycle and increase the probability of user satisfaction and successful system adoption.

Barker's system design method is used to organise the development process in seven fundamental steps: strategy, analysis, design, build, documentation, transition and production. The design lifecycle guided step-by-step the design process in Chapters 7 and 8. The significance of the methods based on the sequential linking between stages, where the outputs of each stage are used as inputs for the next stage, provided a smooth transition between stages in addition to the ability to provide full details and proper documentation for each stage including comprehensive user documentation (user guide) for system use.

8.7.2.3 Evaluation frameworks

The multilevel service evaluation approach is based on principles related to user experience and creating a set of interrelated models that bridge the user experience and design the service offering. The multilevel service design process has four steps: studying the system user experience, designing the service concept, designing the service system, and designing the service encounter. Table 8.2 presented each multilevel service design step with its corresponding stage in the Barker method. Each step is fulfilled with actions to achieve the design stage in the development lifecycle. We guaranteed the involvement of the system user throughout the design and implementation process, which will increase user familiarity with and acceptance of the system.

***The informatics evaluation framework*:** The implementation stage is crucial to the development lifecycle, which confirms the success of the design and attaining the purpose. We therefore sought to apply a health informatics evaluation framework based on five stages of the system development lifecycle and equivalent evaluation level according to the Stead et al. framework [208]. The results of one level of evaluation can be used at a later stage of development or can indicate the need to move to the next level of evaluation. Table 8.1 described in detail the method of evaluation used in the design in each level of system evaluation of the Stead et al. framework. This model was adopted to allow the evaluation process to reflect knowledge in the development process. The use of the informatics framework and the multilevel service design approach has helped to increase the user experience with the system, and constant comments are made in the iterative design lifecycle.

8.8 Contribution and future work

The major contribution of the research is as follows. This thesis highlights current issues faced by Saudi Arabia in health information systems and health informatics and the resulting impact of such issues in critical areas, such as genetic clinics and research centres. In particular, there is a lack of technical solutions to solve the problem of collecting, storing and processing clinical data. This thesis proposed solutions to address these issues in the Saudi context based on the current literature. It then implemented these solutions by designing a novel data management system (G3DMS) for the diagnosis of genetic disorders, which supports the health informatics initiative in Saudi Arabia. The G3DMS has three unique support features: (i) the diagnosis of the genetic condition workflow; (ii) diagnosis decision-making procedures; and (iii) applicability to research studies. Another strength of the G3DMS is that it is applicable to any genetic clinic and genetic research and can be implemented successfully in any low-resource setting.

8.9 Summary

Chapter 7 and Chapter 8 presented a design and implementation of the G3DMS using the Barker method in seven comprehensive design steps, then delivered a data management system that applies to any genetic clinic and research centre in Saudi Arabia. A qualitative evaluation of all stages of the design lifecycle was conducted to support the design with the experience of the system user, thus ensuring user satisfaction to obtain a successful and sustainable application. Therefore, the perspective of the G3DMS is to become a standard system for data collection, management and storage to promote data sharing between genetic institutions. In the future, this system is expected to be part of the infrastructure for a multi-system integration platform for the creation of a Saudi national genetic disease database.

Next, in Chapter 9, the G3DMS is an essential component in the proposed integration platform for integrating genetic disorders data from multiple sources. The integrated data repository of genetic disorders data referred to as GENE2D converts the solution architectural framework presented in Figure 6.1 into a working model.

Chapter 9: A NoSQL Integrated Data Repository of Genetic Disorders Data – GENE2D

9.0 Chapter overview

This chapter presents the design of and implements a "not only SQL" (NoSQL) based integration framework to generate an integrated data repository of genetic disorders data called GENE2D. Section 9.1 introduces the purpose and objectives of this chapter and provides links to previous chapters. Section 9.2 provides an overview of the method followed to achieve the GENE2D. Section 9.3 describes the system architecture of the integration framework and its major components. Section 9.4 presents the complete steps of the design methodology followed to establish an integrated data repository from multiple sources of G3DMS. Section 9.5 presents the results of the integration framework. Section 9.6 focuses on the contribution and future work of the GENE2D. Finally, Section 9.7 summarises this chapter and introduces the conclusion chapter.

9.1 Introduction

This chapter spans the last design and implementation stages, starting from sketching the problem in the abstract world to delivering the complete solution in the real world (Figure 1.1, Figure 1.2). The purpose of this chapter is to produce an integration framework for integrating data from multiple clinics who depend on the G3DMS and answer the research questions, see Figure 1.1, Q6.A: *How can an integration framework be designed for aggregating genetic disorders data from multiple genetic clinics and research centres depending on efficient, cost-effective technologies?* and Q6.B: *How can an integration framework be implemented?* The thesis proposed an architectural framework for the solution in Figure 6.1, including a standalone G3DMS and an integration framework for data from these systems. Chapter 7 and Chapter 8 presented the design and implementation of the G3DMS, respectively, then incorporated an integration framework into the G3DMS. This chapter incorporates an integration framework into the G3DMS to design and implement a NoSQL based on an integrated data repository of genetic disorders data called GENE2D to integrate data from various genetic clinics and research centres in Saudi Arabia and provide an easy-to-use query interface for researchers to conduct their studies on large datasets.

9.2 Method

The physical integration approach is adopted to achieve the intended integrated data repository for GENE2D. The design of the integrated data repository incorporates the extraction, transformation and loading process, which will take place at the source system in G3DMS, which has been developed and implemented in a Saudi genetic clinic. Customised methods were created, specifically for mapping to perform the exchange, transformation and loading process. The NoSQL document database is used instead of the traditional relational model. Therefore, this design benefits from the flexible schema model for a complex data structure as document formats are self-describing, and a collection may include documents with different forms [209]. GENE2D is accessed via a frontend query interface called a Query Builder, explicitly designed for research purposes and for answering ad hoc research questions.

9.3 System architecture

The NoSQL integration framework combines data on genetic diseases from multiple sources into the GENE2D system. The major components involved in the design architecture shown in Figure 9.1 are:

- Integration sources: G3DMS contains genetic disorder data stored in relational tables on a MariaDB server and is the source from which genetic data are extracted, converted and uploaded to the integrated data repository, GENE2D.
- Document store: The NoSQL document store, MongoDB, which uses a collection of documents to store the data as Binary JavaScript Object Notation, is the centralised repository for the integrated data.
- Query interface: The Query Builder interface that provides multiple pages for data retrieval in selective display forms according to the query type selected allows researchers to retrieve content from the database using a custom query to answer simple or complex research questions.

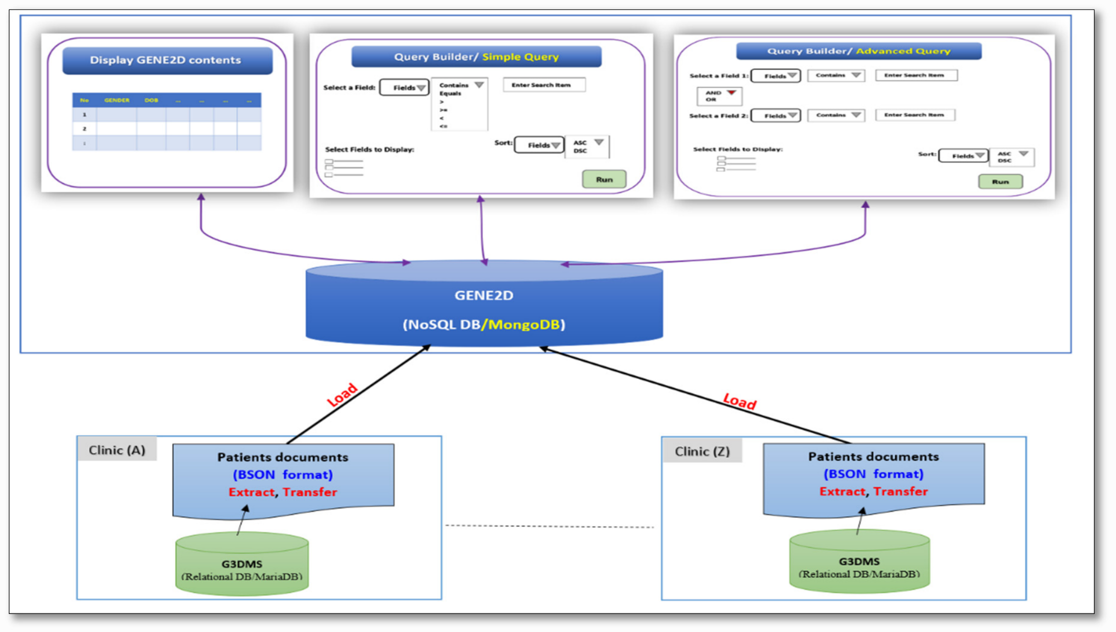


Figure 9.1: NoSQL integrated data repository of genetic disorders data (GENE2D) architecture

9.4 Design methodology

The design methodology of the GENE2D system aims to achieve the final design solution in three fundamental design stages: data acquisition process; data modelling; and system implementation and testing.

The design process involves automated data extraction transformation and load at the source using customised methods. First, the data extraction step results from reading data from the source (G3DMS) using PHP methods for MySQL. Then a physical schema based on the data structure model is developed. Next, the constructed schema for the target system, a NoSQL document database, is used as a model for data transformation using MongoDB/PHP-based methods on the extracted data. After this, patients' data are loaded into individual documents to reside on the Patients collection in the GENE2D system. Last, the GENE2D application interface is implemented, and a validation test is performed.

9.4.1 Data acquisition process

The clinical and genetic data of patients with genetic disorders is available in individual clinics that are collected and managed by the system developed and implemented, called G3DMS [15]. The integration of data from these systems requires a centralised repository to accommodate all data from multiple G3DMSs. Therefore, when acquiring data from different database management systems with various storage structures, the data cannot be directly migrated unless it goes through different processes to be cleaned, organised and converted to other formats and types to adhere to the destination structure. Typically, the integration process includes extracting, transforming and uploading processes, which provide methods for transferring data from the source to the data warehouse database. These operations are carried out in a staging area between the source and destination system where data integrity problems are managed and verified before loading. In this design method, the extracting, transforming and uploading operations are all executed at the source system automatically using customised methods from both world SQL and NoSQL databases. The extracted and converted data are sent from the G3DMS to the GENE2D system with a simple click of the Export button. The data acquisition process starts with extracting data from the G3DMS. The data are cleaned and prepared for transformation to the intended format before uploading to the MongoDB-based GENE2D system.

9.4.1.1 Data extraction

The G3DMS contents range from demographic data (medical record number, genetic number, name, gender, date of birth, nationality, address/city and contact number), clinical characteristics (phenotypes), family history information (inheritance pattern, consanguinity, mother's age, DNA availability, pedigree chart), diagnosis data such as samples, consents, tests, results, photos, research papers, reports, conditions, diagnosis status and treatment plans (plan type, description, date), in addition to physicians' details and contacts. Several criteria are imposed to help refine data selection and extraction from the source.

Extraction criteria

- *Patient-oriented*: The integration framework aims to collect patients' data and create the GENE2D system; therefore, it is a patient-centred database, especially for patients with genetic disorders. Thus, all data extracted should be related directly to the patient, so any other data, such as physician information, should not be included.
- *Research requirements*: Only data that are useful for research related to genetic disorders should be included. Unnecessary data, such as patient treatment plans and contact details, will not be considered. Also, only cases of "confirmed" diagnoses of genetic conditions should be transferred. Therefore, cases with "provisional" diagnoses are postponed until the diagnosis status changes to "confirmed". Any incomplete patient data which misses important information is not extracted for data quality purposes.

- *Health Insurance Portability and Accountability Act compliant*: The GENE2D system contents will be available online to researchers, so they must adhere to specific rules and regulations related to the use of identifiable personal health information [36]. Since the dataset should not include an identifier which can be used to link back to the patient, any information which may lead to the identity of a patient will not be moved from the G3DMS. This includes names, photos, samples and family pedigree charts or any documents containing patient identification information. However, de-identification can be achieved by masking or encoding a unique key such as the medical record number to solve the problem of updating the data in the GENE2D system while preventing duplicate entry.

Extraction method

The G3DMS uses the relational database management system MariaDB server as a backend database for storing data related to patients with genetic conditions. The extraction process is done by applying the following steps. First, a connection to the source G3DMS database is created using PHP-function MySQLi_Connect. Next, the database is navigated using the primary key medical record number and making the appropriate linkage to access all data distributed in various tables. Then, each criterion is applied as a condition to the query in the related table to extract only the required fields; for example, the diagnosis status is confirmed.

```
$Sql2 = "SELECT `CONDITION_NAME` as conditions FROM `diagnosis` WHERE `MRN`=$MRN AND
`STATUS`='Confirmed'".
```

Similarly, some required patient data are retrieved using joins between tables such as patient's phenotypes and clinical characteristics, and tests and results. Then, the query results are retrieved and accessed to obtain each value and store it in an array of its type. Finally, all the resulting arrays are merged in one single array after checking empty values to ensure that incomplete data is excluded. Now all patient data stored in one array are ready to be converted to the specified format and data type before passing it to the writing method in MongoDB.

Extracted data

Table 9.1 lists the data that meet the extraction criteria and are eligible to be copied from the G3DMS to the GENE2D system. The list includes field names, data types and role in terms of their relationship with the primary entity or object "Patient". All data fields will be moved without any changes except the primary key, the medical record number, which has to go through encoding for de-identification purposes and the hashing function is used to produce a new unique ID for each patient in the organisation.

Table 9.1: List of fields extracted from the G3DMS (genetic disorders diagnosis data management system)

Field name	Type	Relationship
Masked (MRN) → _id	Primary key	Patient’s proprietary
Gender	Attribute	Patient’s proprietary
Dob	Attribute	Patient’s proprietary
City	Attribute	Patient’s proprietary
Nationality	Attribute	Patient's proprietary
Consanguinity	Attribute	Patient’s proprietary
InheritancePattern	Attribute	Patient’s proprietary
MotherAge	Attribute	Patient’s proprietary
Phenotypes [pheno1, pheno2, ...]	Entity	One-to-many
Conditions [cond1, cond2, ...]	Entity	One-to-many
Tests [{test 1, result 1}, {test 2, result 2}, ...]	Entity	One-to-many

9.4.2 Data modelling

Document-oriented data models deal with documents, so the storage structure is based on the primary unit of collections (tables as in the relational model) made of individual documents (tuples as in the relational model). Documents, in general, contain all the information about the entity collectively, where each document stores a JSON object in the form of key-value pairs, although the document store does not require data modelling where there is no schema definition obligation. The design methodology supports the data modelling process to generate a schema to define the plan for data organisation for better query performance and quicker response time. Therefore, defining the schema as a proactive step before transferring the data assists the semantic transformation and increases the overall performance of the application.

Modelling in document databases is as essential as in relational modelling; there is still a schema, but it is not enforced as in relational modelling, although the design of the schema or the development of the data model for a document-oriented database does not follow a standard fixed method such as the relational modelling approach (conceptual, logical, physical) models. A useful data model will result in a robust schema that optimises query performance while preserving time and storage efficiency. Also, balancing application requirements, database engine performance characteristics and data retrieval patterns are a significant challenge for data modelling. Therefore, application usage must be considered, such as queries and updates, as well as the data structure and organisation within the database. Consequently, the purpose of the data modelling process is to produce a useful data model that supports the workflow in the application of the GENE2D system, which aims to retrieve and answer any research question related to patients with genetic diseases. Therefore, the focus is on the read performance being fast and efficient using dynamic querying options.

9.4.2.1 Schema plan

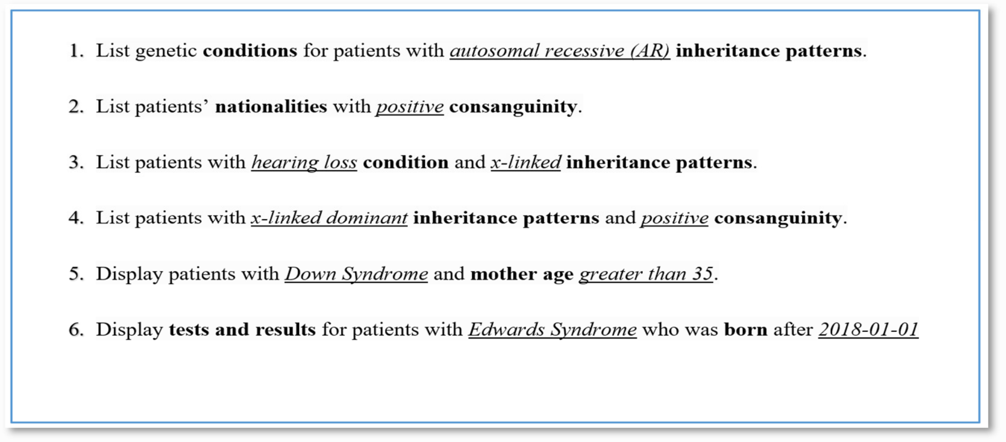
The most critical step when developing the data model is the decision on the way entities (objects) should reside in the database. How data are accessed and queried will determine how the data are stored, so data retrieved together should be stored collectively. Since the priority is to increase query performance during the read and update operations, the schema design plan considers minimising the need for the database to make any joins and reduces Input/Output operations. Hence, the design decision of a patient-oriented database for the GENE2D system is to store all the datasets in one single collection which will encapsulate all the patients' documents; each patient record will be stored as an independent document. Next, the way the patient's data are organised inside the document (the physical model) is determined by the data modelling method adopted to develop the physical schema. Two design guidelines are followed for schema development. First, the data dependencies and relationships between the attributes and entities of the extracted data from the G3DMS are analysed, as listed in Table 9.1. Second, the application-specific access patterns for the GENE2D system are identified, i.e., the query patterns to be supported.

Data dependencies and relationships

The patient-oriented design focus on the GENE2D system mandates that all attributes and entities in the database are directly related to the patient entity. The patient's proprietary attributes, such as medical record number, gender, date of birth, nationality, city, consanguinity, inheritance pattern and mother's age, are totally dependent on the patient entity with a relationship of one-to-one. Entities such as conditions, phenotypes, and tests and results also show dependency on the patient's entity but with a different relationship of one-to-many.

Query patterns

The purpose of the GENE2D system is to apply it to research studies, specifically to help answer research questions related to diagnostic information for Saudi patients with genetic disorders. Defining the type of queries required by research studies in the area of genetic diseases, to answer *What question to expect?* will determine the data access patterns and the application workflows. Also, it will help in the identification of primary keys, indexes, denormalised attributes and their organisation within the document. Figure 9.2 displays a sample of use cases for questions that can be expected to conduct research in the field related to the diagnostic information on patients with genetic conditions in Saudi Arabia. The central entity in the database appears to be the "patient" object and its proprietary attributes that are repeatedly present in all queries. Therefore, patient_id is considered as the primary key, and all its attributes need to be denormalised and considered for possible indexing. Also, entities that have shown dependence on the "patient" entity, such as conditions, phenotypes, and tests and results, are also subject to indexing.



1. List genetic **conditions** for patients with *autosomal recessive (AR)* **inheritance patterns**.
2. List patients' **nationalities** with *positive* **consanguinity**.
3. List patients with *hearing loss* **condition** and *x-linked* **inheritance patterns**.
4. List patients with *x-linked dominant* **inheritance patterns** and *positive* **consanguinity**.
5. Display patients with *Down Syndrome* and **mother age** *greater than 35*.
6. Display **tests and results** for patients with *Edwards Syndrome* who was **born** after *2018-01-01*

Figure 9.2: Query patterns represented in use cases

9.4.2.2 Schema design

Dependencies, relationships between entities and attributes, and query patterns are the key elements in the design of the physical schema. The design focuses on document structure and the representation of data relationships. The results of a discussion of the schema plan and the design guidelines support the choice to store all patient-related data in one collection.

Database structure

Figure 9.3 presents the general database structure for the GENE2D system in MongoDB, namely the database container "GENE2D", the collection "Patients" and the documents where each document denotes patient information represented by the structure of a JSON object. Each patient document or JSON object contains many fields such as id, gender and date of birth. The decision is to aggregate all data needed to process any query in one place so that multiple queries will access the same data in a different combination.

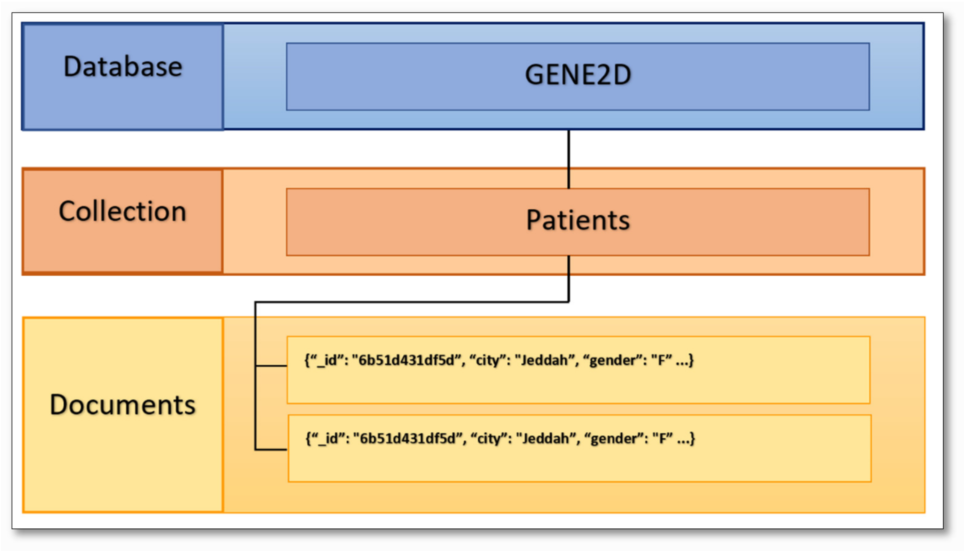
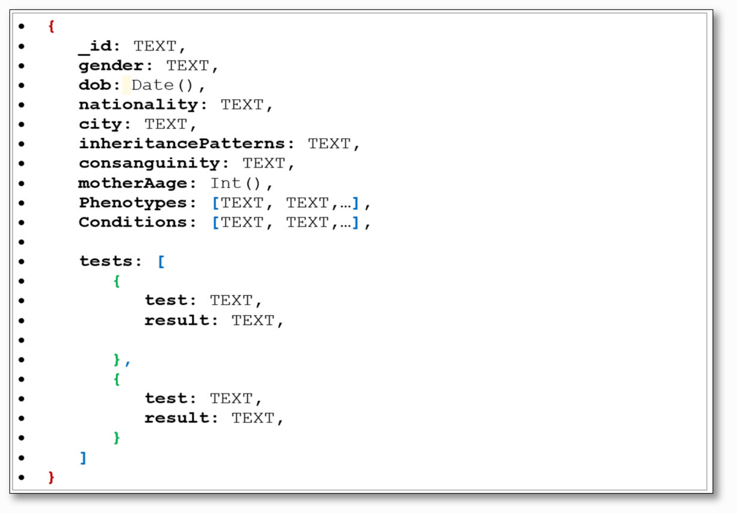

Database
GENE2D
Collection
Patients
Documents
{“_id”: "6b51d431df5d”, “city”: "Jeddah”, “gender”: "F” ...}
{“_id”: "6b51d431df5d”, “city”: "Jeddah”, “gender”: "F” ...}

Figure 9.3: GENE2D database structure

I. Embedded data model

The aim is to store the data in a query-friendly structure to facilitate query operations. Therefore, the denormalised or embedded data model to represent the data structure and constitute the physical schema will enable the related data to be stored in a single document which eliminates the need for joins. Embedding refers to the insertion of an individual document within another related element (document) which is frequently used together, i.e., nesting one document into another. First, some of the extracted patients’ properties such as _id, gender, date of birth, nationality, city, inheritance patterns, consanguinity and mother’s age, will be included in the “patient” object as flat list attributes (fields), for example {“_id”: “6b51d431df5d”, “city”: “Jeddah”, “gender”: “F”...}. Entities include phenotypes and conditions, of one-to-many relationships, the size of the list is limited. Therefore, the decision is to embed both entities as arrays in the “patient” object, for example, {“phenotypes”: [“ph1”, “ph2”,…],…}. Regarding the test entity which consists of a list of {test, result}, there is an option to embed test data as an array of objects inside the patient document or to reference the “test_id” inside the patient document and store the test data in a separate document. We embed and reduce the joins because the tests are bounded with limited growth in size, for example, { “tests”: [{ “test1”, “result2”}, {test2,result2},…], …}. Figure 9.4 shows the general design of the physical schema, which represents the patient’s document as a JSON object. The overall decision to embed corresponds with the purpose of the application function as a query builder interface.



```
{
    _id: TEXT,
    gender: TEXT,
    dob: Date(),
    nationality: TEXT,
    city: TEXT,
    inheritancePatterns: TEXT,
    consanguinity: TEXT,
    motherAage: Int(),
    Phenotypes: [TEXT, TEXT,…],
    Conditions: [TEXT, TEXT,…],

    tests: [
       {
          test: TEXT,
          result: TEXT,

       },
       {
          test: TEXT,
          result: TEXT,
       }
    ]
}
```

Figure 9.4: Schema design for the patient document

9.4.2.3 Transform

Moving data from a relational database management system to a NoSQL document store such as MongoDB requires a carefully planned transformation process based on the data structure on both the source and target databases. The schema design is the key for database construction to meet the requirements of the GENE2D application workflow. The transformation process depends on the schema design presented in Figure 9.4, which is used as a guide for mapping and transforming the extracted data from the source G3DMS to be loaded and stored in the MongoDB destination. Customised methods are implemented to map the extracted data to fall into the schema design structure according to the defined mapping guidelines.

Mapping guidelines

Setting guidelines for data transformation allows us to define and restructure the extracted data using customised operations, for example, for data cleaning and conversion of units such as date and time before loading into the target MongoDB-based GENE2D system.

Support aggregation and a void join operation

Adopting the embedding mechanism will allow all patient data to be accessed together. Therefore, using an array to concatenate all patient data to be mapped in a single operation to a JSON object adheres to the embedding model.

Field naming

The field names in the G3DMS source are consistent for use in research, so expressions are used to name the fields so they can be identified by researchers. Therefore, most field names are mapped directly without changes, whereas slight modifications were made to others, so they are more specific but have the same meaning as (clinical_description → phenotypes), and so on.

Multivalued fields

In some cases, a patient may have a list of phenotypes, conditions and tests. These multivalued fields need to be stored in a way that they can be retrieved correctly for the individual patient as well as for the cohort. While reading data from the related joint tables of a specific patient using the primary key medical record number, we use a multidimensional associative array to store the field name in the key such as array [phenotypes] and a numeric array for storing the values, for example: Array ([phenotypes] => Array ([0] => pheno1 [1] => pheno2 [2] => pheno3)).#For the tests array, which includes sets of {test, result}, the multidimensional associative array will look like this:

Array ([tests] => Array ([0] => Array ([test] => test1 [result] => result1) [1] => Array ([test] => test2 [result] => result2))). Therefore, when mapping to the MongoDB document, the array [key] of the associative array [phenotypes] will be used as the field name for the array in the document and the same for the conditions and test arrays.

Transforming methods

I. De-identification

The patient identifier field, the medical record number, is the most critical value and so it needs to be treated differently, as it is the conventional method of identifying a patient in Saudi hospitals and clinics. Therefore, the selection of this field to be encoded automatically before sending the data to the GENE2D system is very significant to preserve patients' confidentiality and prevent duplication of data during the update processes. The design uses a one-way hash algorithm on the medical record number field to anonymise the patient's data while allowing researchers to update a specific record with a modification. The standard hash algorithm is implemented to transform the medical record number string into another string "_id" in a way that no other operation can retrieve the original medical record number. $_id_array['_id']= hash("sha256",$MRN, $raw_output = FALSE)> The "_id" will be assigned as the primary key that uniquely identifies the document.

II. Indexing

When uploading data to MongoDB, each document is indexed automatically using the unique document "_id" as a default primary key set by MongoDB/ObjectId class. But in our case, we assigned the masked medical record number as the document primary key. So, the "_id" is set as the index for the document. Indexing on all fields is also valuable in that it makes the reading process more manageable, the query operation much faster and obtains more precise results by reaching all the data inside the arrays and object arrays. One of the significant advantages of MongoDB compared to other document stores is that it allows any field in the document to be set as an index. Therefore, the indexing function createIndex([‘$**’ => 1]) is used to create a wildcard index on all the fields and subfields in a document.

III. Convert field types

Almost all the fields with string type are transformed without any changes, but fields like date type “dob” and integer type “motherAge” must be validated and converted to the appropriate MongoDB format to be valid for a query of its type. For the integer field, a simple PHP (int) function is used to force the type before loading. The “dob” date string in the source database must be converted to conform to the MongoDB format which uses the BSON date format. So, to convert the date, first, the PHP function strtotime() is used to map the date string (y-m-d) into a Unix timestamp. Then, the resulting timestamp is passed to MongoDB to be converted and stored as a BSON date.

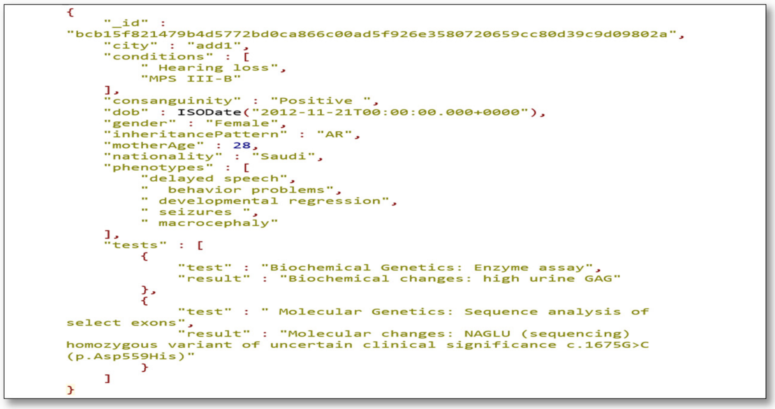
9.4.2.4 Load

I. Loading methods for patient data

The stage of loading the transformed data is preceded by making a connection to a MongoDB server. In the design, PHP functions are used to read from the source in a MariaDB server and write to the MongoDB server using MongoDB classes from the MongoDB library. We created a new connection using the MongoDB localhost at port 27017:# $connection=newMongoDB\Driver\Manager('mongodb://localhost:27017')#Then we used a new class from MongoDB\Driver\BulkWrite, which can be constructed with (insert or update) operations to store the patient’s array contents as a JSON object (document). For the bulk write we use the update operation with “upsert” option instead of using insertMany() method to allow multiple updates to the same patient document without having the problem of duplicate document error due to “_id” existence in the collection. The insert method will work the first time; however, it will fail to write to the collection if there is a duplicate “_id” for an existing document. Therefore, we used the update method as follows: first, create new object for bulkWrite, $docs = new MongoDB\Driver\BulkWrite; then the update method, $docs->update(['_id'=> $_id_array['_id']],['$set'=> $patient_array], ['multi'=>false, 'upsert' => true]);#finally, after storing all patient array contents in the bulkWrite object, the executeBulkWrite was used to pass the patient data to create a new document for new patients or update an existing patient document with the new data as follows: $result= $connection->executeBulkWrite('GENE2D.Patients', $docs).# The executeBulkWrite class creates a new database if it does not exist, so the database “GENE2D” and the collection “Patients” will be created for the first time; otherwise, it just updates the existing collection with the new entry using the update operation. Confirmation messages are displayed to verify the write operation using the following methods: $result->getMatchedCount(), $result->getModifiedCount(),#and#$result->getUpsertedCount(). Figure 9.5 demonstrates the structure and content of the resulting entry of one patient document, which contains the hashed “_id” as a primary key and arrays such as conditions and phenotypes with their contents list, and also the tests array with its content object set of {test, result}, in addition to other patient’s data of string type, and “dob” in an ISODate format, and “MotherAge” as an integer of 32-bit.

II. Loading method for user’s login data

The user’s login data is also mapped from the G3DMS database to a separate collection “login” in the GENE2D database. A new “user_id” was constructed by customising a function to combine the username and the hashed password in one string. Then MongoDB bulkWrite method was used to write users’ login data into the “login” collection. The objective of moving users’ accounts from the local system G3DMS to the integrated GENE2D is to provide immediate access to larger datasets for researchers in the participating clinics and to encourage other centres and clinics to participate in the integration. #



```
{
    "_id" :
"bcb15f821479b4d5772bd0ca866c00ad5f926e3580720659cc80d39c9d09802a",
    "city" : "add1",
    "conditions" : [
        " Hearing loss",
        "MPS III-B"
    ],
    "consanguinity" : "Positive ",
    "dob" : ISODate("2012-11-21T00:00:00.000+0000"),
    "gender" : "Female",
    "inheritancePattern" : "AR",
    "motherAge" : 28,
    "nationality" : "Saudi",
    "phenotypes" : [
        "delayed speech",
        "  behavior problems",
        " developmental regression",
        " seizures ",
        " macrocephaly"
    ],
    "tests" : [
        {
            "test" : "Biochemical Genetics: Enzyme assay",
            "result" : "Biochemical changes: high urine GAG"
        },
        {
            "test" : " Molecular Genetics: Sequence analysis of
select exons",
            "result" : "Molecular changes: NAGLU (sequencing)
homozygous variant of uncertain clinical significance c.1675G>C
(p.Asp559His)"
        }
    ]
}
```

Figure 9.5: Training data for a patient document

9.4.3 System implementation and testing

9.4.3.1 Development environment

Since we used the XAMPP stack which contains Apache webserver, MySQL and PHP compiler for the development of G3DMS, we find it useful to set up MongoDB along with PHP in the same stack. XAMPP stack will allow us to work with both databases (the source system MySQL MariaDB and the target system NoSQL MongoDB) in the same environment.

First, we installed MongoDB 4.2 for windows, with MongoDB Compass GUI to visualise the data and check the entry. Since PHP does not communicate directly to the MongoDB server, we installed MongoDB-PHP driver for Windows following the next steps [210]:

i. Download the MongoDB-PHP driver
ii. Add 'php_mongo.dll' file to the PHP extension directory.
iii. Update 'php.ini' file and add 'extension=php_mongo.dll' to it
iv. Restart web server (Apache server)

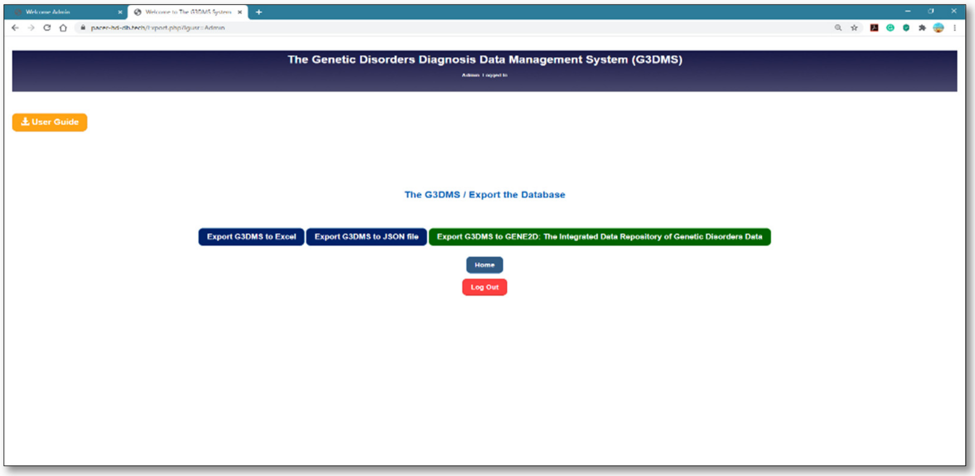
Also, we installed the PHP Library for MongoDB with Composer [211], and store the vendor folder in the same directory with the PHP files to be easily accessed by the application interface files.

9.4.3.2 Application interface

We developed our web pages for the system user interface in the XAMPP platform using the Apache server to run and test the web interface before publishing the GENE2D system on a live server. Appendix 9.1 provides all the source code for GENE2D.

I. Update the G3DMS application interface

The three main steps in the integration process of data extraction, conversion and load are combined into one operation at the source G3DMS. To implement the process, we included a button to export the data from the G3DMS to the GENE2D system in the G3DMS application interface, as shown in Figure 9.6. A PHP file called "export_to_MongoDB.php" was developed in the application files of the G3DMS to read data from the MySQL database tables and then it is converted using classes from the MongoDB Library for PHP.



The Genetic Disorders Diagnosis Data Management System (G3DMS)

The G3DMS / Export the Database

Welcome Admin
Welcome to The G3DMS System
Admin I ogged In
User Guide
Export G3DMS to Excel
Export G3DMS to JSON file
Export G3DMS to GENE2D: The Integrated Data Repository of Genetic Disorders Data
Home
Log Out

Figure 9.6: Export page from the G3DMS to the GENE2D system

II. GENE2D application interface development

The primary goal of the system is to provide an easy-to-use user interface taking into consideration novice users who may be professionals in their fields but who do not have to learn the query language to conduct their search. Therefore, the user interface design focuses on connecting the system to a user action using clear messages, as well as looking at the simplicity of the user interface and visual aspects such as clarity and colour. Since the system will be used by researchers from across Saudi Arabia, officials in the sector will determine the validity of the entry and the use of the system, but as a first step, access will be granted to users of the G3DMS. The application consists of several web pages for Login, Reset Password, Query Builder Home, Display Database Content, Simple Query Builder and the Advanced **Query Builder.**

Authentication pages

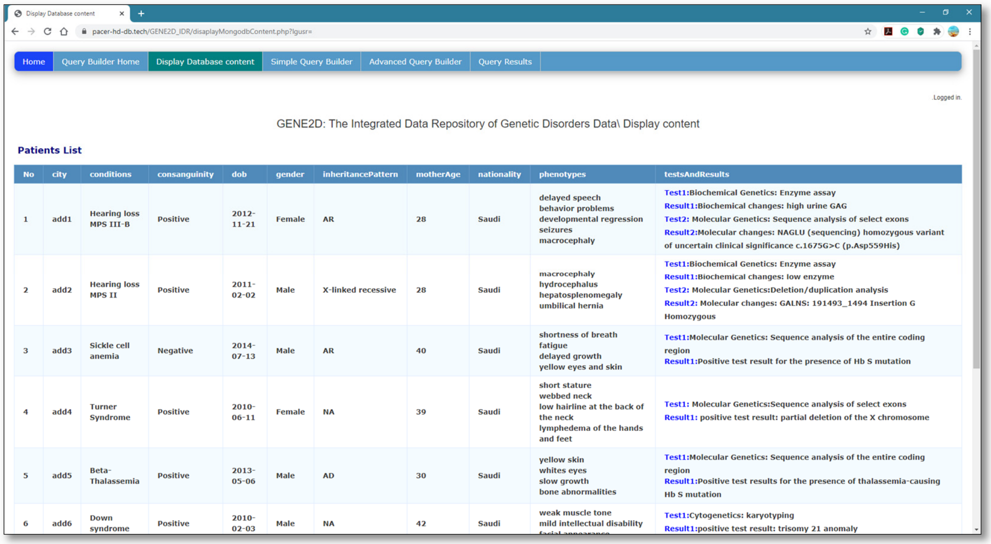
Since the system will be accessed by an authorised researcher using the same username and password for the G3DMS, the first page of the application is a login authentication page that verifies the user's username and password before they are allowed to proceed to the system. The username and password used for the G3DMS authentication are the same. All users' login data are transferred along with the patient data to be stored in a separate collection called "login" in the GENE2D database. Also, the system allows the user to change their password in the GENE2D system through the Reset Password page, which updates the user information in the "login" collection with the new password.

Query Builder home page

This page allows the user to select one of three options to navigate the database: the Display Database Content page to display the entire database, the Simple Query Builder page which relies on one field of the document to answer questions, and the Advanced Query Builder page to query two fields using a logical operator (and, or) and to allow the result to be sorted (ASC or DESC) by any field.

Display Database Content page

This page creates a new connection to the MongoDB server and reaches the "GENE2D" database and the "Patients" collection. As shown in Figure 9.7, the resulting data are presented using the HTML table format to display the output in a tabular form.



GENE2D: The Integrated Data Repository of Genetic Disorders Data\ Display content

Patients List

No	city	conditions	consanguinity	dob	gender	inheritancePattern	motherAge	nationality	phenotypes	testsAndResults
1	add1	Hearing loss MPS III-B	Positive	2012-11-21	Female	AR	28	Saudi	delayed speech behavior problems developmental regression seizures macrocephaly	Test1:Biochemical Genetics: Enzyme assay Result1:Biochemical changes: high urine GAG Test2: Molecular Genetics: Sequence analysis of select exons Result2:Molecular changes: NAGLU (sequencing) homozygous variant of uncertain clinical significance c.1675G>C (p.Asp559His)
2	add2	Hearing loss MPS II	Positive	2011-02-02	Male	X-linked recessive	28	Saudi	macrocephaly hydrocephalus hepatosplenomegaly umbilical hernia	Test1:Biochemical Genetics: Enzyme assay Result1:Biochemical changes: low enzyme Test2: Molecular Genetics:Deletion/duplication analysis Result2: Molecular changes: GALNS: 191493_1494 Insertion G Homozygous
3	add3	Sickle cell anemia	Negative	2014-07-13	Male	AR	40	Saudi	shortness of breath fatigue delayed growth yellow eyes and skin	Test1:Molecular Genetics: Sequence analysis of the entire coding region Result1:Positive test result for the presence of Hb S mutation
4	add4	Turner Syndrome	Positive	2010-06-11	Female	NA	39	Saudi	short stature webbed neck low hairline at the back of the neck lymphedema of the hands and feet	Test1: Molecular Genetics:Sequence analysis of select exons Result1: positive test result: partial deletion of the X chromosome
5	add5	Beta-Thalassemia	Positive	2013-05-06	Male	AD	30	Saudi	yellow skin whites eyes slow growth bone abnormalities	Test1:Molecular Genetics: Sequence analysis of the entire coding region Result1:Positive test results for the presence of thalassemia-causing Hb S mutation
6	add6	Down syndrome	Positive	2010-02-03	Male	NA	42	Saudi	weak muscle tone mild intellectual disability facial appearance	Test1:Cytogenetics: karyotyping Result1:positive test result: trisomy 21 anomaly

Figure 9.7: Display of the database content page

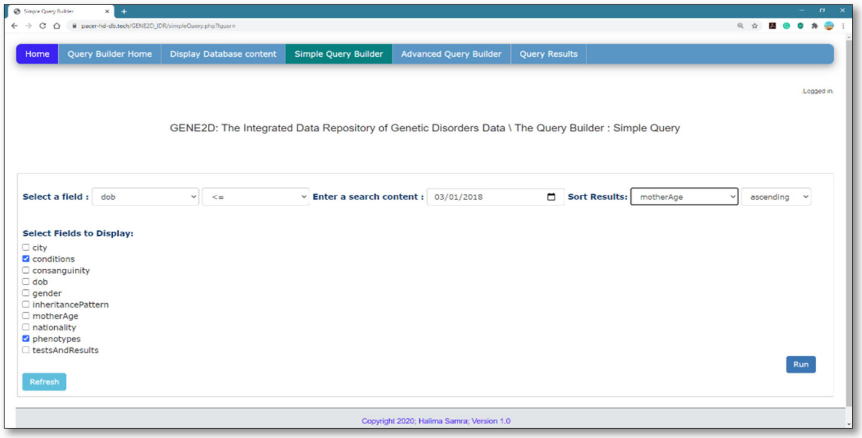
Simple Query Builder page

This page allows the user to select a field from the document to search, specify the search criteria, enter search contents, select a field to sort the results and tick all the extra fields of information to display beside the essential selection, as presented in Figure 9.8.

***Select a field to query*:** Select an option from the drop-down menu, which displays all fields in the documents. This menu changes dynamically according to the existing fields in the documents. This provides flexibility and guarantees data availability. The user selection from this menu affects the option that appears in the next menu of the search criteria. For example, if the user chooses to query on the field “conditions”, then the adjacent menu will display options (contains and equals), and the text box will allow the user to type a search item. But for the option “consanguinity”, the next menu will change to show options (positive or negative). If “dob” is selected as the search criteria, it will allow a list that matches with the data type, therefore the menu of operators (equals, >,>=,<,<=) appears accordingly; also, a date type insertion box surfaces to allow the selection from the date calendar. Similarly, for the numeric “motherAge” selection, the same search operator shows an input box for the numeric. When selecting the option of tests and results, the next menu pops up with two options (test or result) and a fixed search criterion (contains) and a text box to enter the search phrase. We used jQuery functions to control the appearance of menus on the page using hide( ) and show( ) methods according to the user selections. Also, the page is supported by a “Refresh” button to reload the page as well as messages to verify the input fields when the user clicks the “Run” button.

Sort results: The same list of current fields appears to select one to sort the query result accordingly with an option to specify the sort order (ascending or descending) that will apply to all data types.

Fields to display: Figure 9.8 shows a list of all field names from the patient document which are displayed next to checkboxes from which the user can choose. Selecting from this list enables the display of additional information as needed to support the research question. This increases the accuracy of the query results and displays the answer from multiple dimensions and perspectives. A verification message appears to remind the user to select at least one field to display if the selection is not made when the query is submitted.



GENE2D: The Integrated Data Repository of Genetic Disorders Data \ The Query Builder : Simple Query

Select a field : dob <= Enter a search content : 03/01/2018 Sort Results: motherAge ascending

Select Fields to Display:
☐ city
☑ conditions
☐ consanguinity
☐ dob
☐ gender
☐ inheritancePattern
☐ motherAge
☐ nationality
☑ phenotypes
☐ testsAndResults

Copyright 2020; Halima Samra; Version 1.0

Figure 9.8: Simple Query Builder page

Simple Query Results page

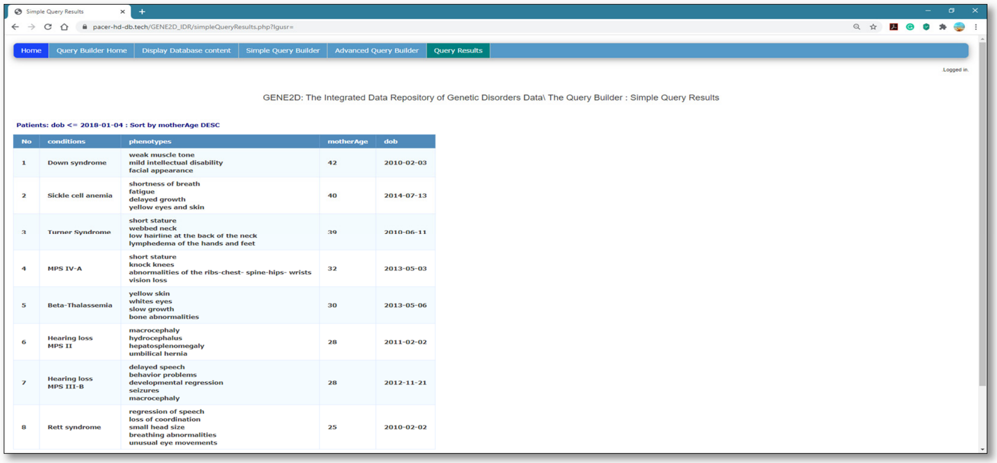
This page will be called as a response to the submit button "Run" in the Simple Query Builder page. All selections will be posted to be managed individually using PHP to get the results from MongoDB and display the query result as output in HTML table format. First, a connection was made to the MongoDB server, then the "GENE2D" database and "Patients" collection were invoked. We used some MongoDB BSON type classes from the library to help perform actions on specific fields such as converting date and matching regular expressions (MongoDB\BSON\UTCDateTime) and (MongoDB\BSON\Regex).

The query format: Each option posted is directed to the designated section to conduct the appropriate query, depending on the search criteria (contains, equals, >, <, etc). For example, in the case of selecting the search criterion "contains", we use the query form with the regular expression pattern matching as follows: $query = [trim($searchField) => new Regex($searchItem, 'i')];. However, for "equals", a direct match will be used as follows: $query = [trim($searchField) => trim($searchItem)];# But, in case of using operators (> "$gt", >= "$gte", < "$lt", and <= "$lte") the

integer value such as "motherAge" will be defined as follows: $query = [trim($searchField) => [$searchCriteria => intVal($searchInt)]];also, the BSON date type fields will be queried using the following format: $query=[trim($searchField)=>new MongoDB\BSON\UTCDateTime(strtotime((string)$searchDate."24:00:00")*1000)];.
On the other hand, the "tests" array of objects {test, result}, which have a different structure of storage, will be queried differently. For example, when selecting the search field as "testsAndResults" and the specify "test" or "result" to query, the query format will be as follows: $query=[trim($searchField) => ['$elemMatch' => [$testsObject => new Regex($searchItem, 'i')]]]>

The sort option: The sort field and sort order will be applied as an option to be added to the find(query, options) method, so the format for the sort option will be as follows: $options = ['sort' => [$sortField => $sortOrder]].

The query result: The MongoDB find() method retrieves the result from the collection using two parameters of the query array and options: $cursor= $collection->find($query, $options). The cursor object will be iterated using foreach method to loop through all the resulting documents values and display only the selected fields. The results will be organised in a table format with a dynamic table header that reads from the selected fields and user input. The count($query) method, which returns the number of documents retrieved as a result of executing the query is used to display the total number of patients. Some query results, such as the date, must be set to a human-readable format; thus, the MongoDB to DateTime( ) function is used for this purpose, and then the format function format ('Y-m-d') is used to display the date. Figure 9.9 shows the results of the research question presented in Figure 9.8: display conditions and phenotypes for patients who were born in 03/01/2018 and after, and sort the results for mother's age in descending order.



GENE2D: The Integrated Data Repository of Genetic Disorders Data\ The Query Builder : Simple Query Results

Patients: dob <= 2018-01-04 : Sort by motherAge DESC

No	conditions	phenotypes	motherAge	dob
1	Down syndrome	weak muscle tone mild intellectual disability facial appearance	42	2010-02-03
2	Sickle cell anemia	shortness of breath fatigue delayed growth yellow eyes and skin	40	2014-07-13
3	Turner Syndrome	short stature webbed neck low hairline at the back of the neck lymphedema of the hands and feet	39	2010-06-11
4	MPS IV-A	short stature knock knees abnormalities of the ribs-chest- spine-hips- wrists vision loss	32	2013-05-03
5	Beta-Thalassemia	yellow skin whites eyes slow growth bone abnormalities	30	2013-05-06
6	Hearing loss MPS II	macrocephaly hydrocephalus hepatosplenomegaly umbilical hernia	28	2011-02-02
7	Hearing loss MPS III-B	delayed speech behavior problems developmental regression seizures macrocephaly	28	2012-11-21
8	Rett syndrome	regression of speech loss of coordination small head size breathing abnormalities unusual eye movements	25	2010-02-02

Figure 9.9: Simple Query results page

Advanced Query Builder page

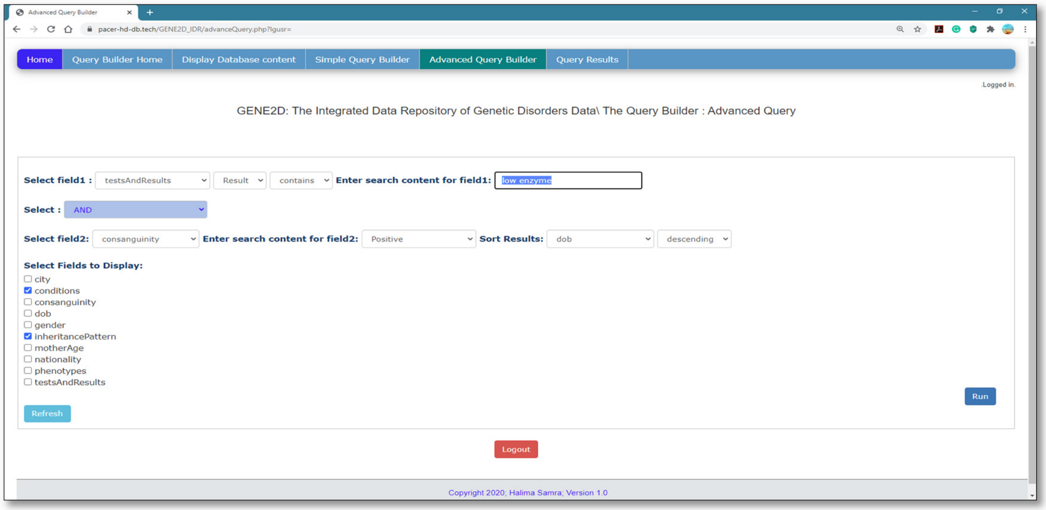
Figure 9.10 shows the advanced query page with options for a complex query, and it may include two fields for direct participation for each field in a separate query with the option to link the two queries with the logical operator (and/or) to match the result. Furthermore, similar to the simple query, any field can be used to sort the results as well as display additional information. Also, the results are presented in the form of a table.

Select fields to query: This page provides two drop-down menus to select a field to participate in two queries and an options list to choose search criteria and input search elements for each query. Note that the options fields are independent as the user can query in the first list of fields or the second list separately or use both lists.

Logical operator: There is an additional list of options to select a logical operator (and, or) to match the results from both queries.

Sort results: It allows the results to be sorted according to one of the fields marked in either ascending or descending order.

Fields to display: Additional information can be provided with the query result by selecting more fields to show from the list as depicted in Figure 9.10.



GENE2D: The Integrated Data Repository of Genetic Disorders Data\ The Query Builder : Advanced Query

Select field1 : testsAndResults | Result | contains **Enter search content for field1:** low enzyme

Select : AND

Select field2: consanguinity **Enter search content for field2:** Positive **Sort Results:** dob | descending

Select Fields to Display:

☐ city
☑ conditions
☐ consanguinity
☐ dob
☐ gender
☑ inheritancePattern
☐ motherAge
☐ nationality
☐ phenotypes
☐ testsAndResults

Copyright 2020; Halima Samra; Version 1.0

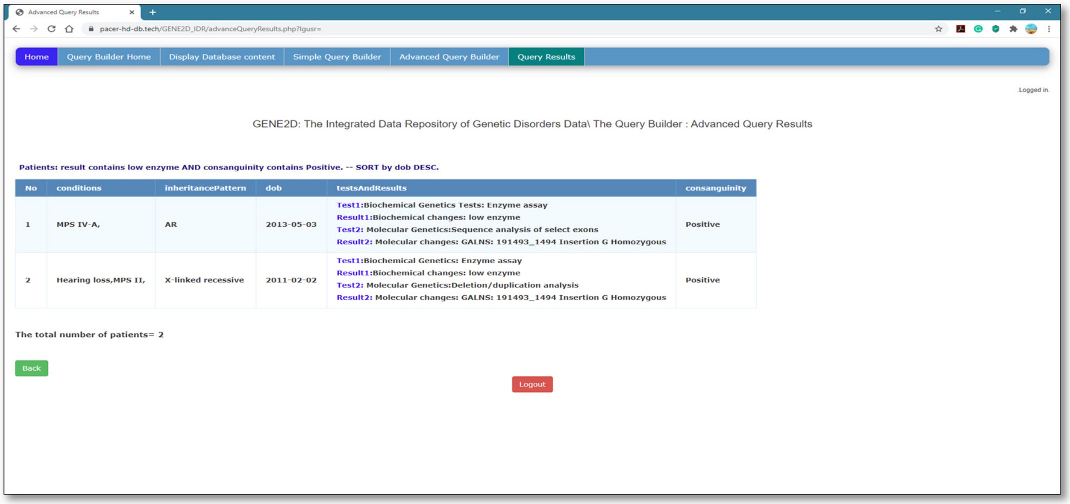
Figure 9.10: Advanced Query Builder page

Advanced Query Results page

The response on this page takes different paths to perform operations, depending on the options specified in the Advanced Query Builder page. First, if the user decides to use only one field, then the page directs the selection to its specific function to execute the query as it did on the Simple Query Results page if both fields are selected and a logical operator, as well as the other related option for both fields. Then the code directs the selection and entry which is posted from the previous page to be executed depending on the search criteria and the type of fields selected as explained in detail in the Simple Query Results page. Here is an example of using two queries with a logical operator:

```
$query = [   $lgOperation => [   [ trim($searchField) => new Regex($searchItem, 'i')   ],   [ trim($searchField2) => new Regex($searchItem2, 'i')   ]   ]   ];
```

Figure 9.10 presents the research question: display the conditions and inheritance patterns for patients with positive consanguinity, and their test results contain "Karyotype". Figure 9.11 shows the Advanced Query Results page that displays the results of a complex query presented in Figure 9.10. The results are organised in a table format with the count of the number of documents retrieved.



GENE2D: The Integrated Data Repository of Genetic Disorders Data\ The Query Builder : Advanced Query Results

Patients: result contains low enzyme AND consanguinity contains Positive. -- SORT by dob DESC.

No	conditions	inheritancePattern	dob	testsAndResults	consanguinity
1	MPS IV-A,	AR	2013-05-03	Test1:Biochemical Genetics Tests: Enzyme assay Result1:Biochemical changes: low enzyme Test2: Molecular Genetics:Sequence analysis of select exons Result2: Molecular changes: GALNS: 191493_1494 Insertion G Homozygous	Positive
2	Hearing loss,MPS II,	X-linked recessive	2011-02-02	Test1:Biochemical Genetics: Enzyme assay Result1:Biochemical changes: low enzyme Test2: Molecular Genetics:Deletion/duplication analysis Result2: Molecular changes: GALNS: 191493_1494 Insertion G Homozygous	Positive

The total number of patients= 2

Figure 9.11: Advanced Query Results page

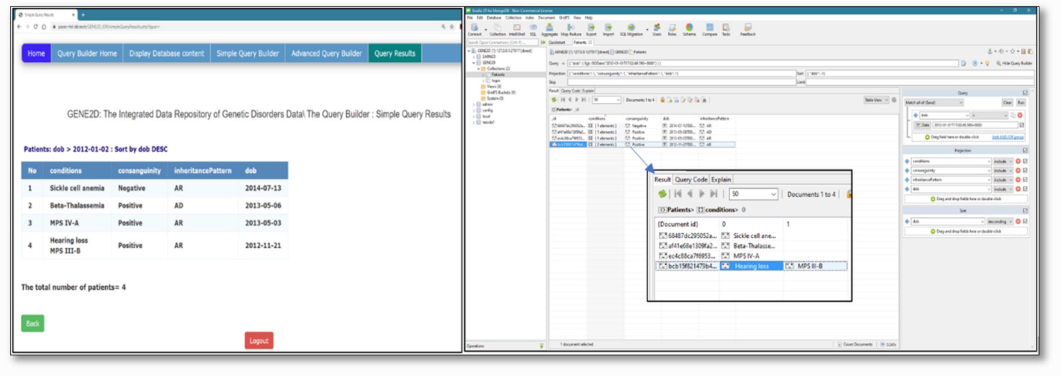
9.4.3.3 Validation test

I. Query testing

We downloaded an open-source non-commercial version of Studio 3T for MongoDB. Studio 3T, one of the most popular MongoDB Graphical User Interface tools, provides a visual editor to write and edit queries. This tool is used to connect to the same localhost of MongoDB, where the GENE2D database resides. The same query is implemented in the GENE2D system, and the Studio 3T and the results are compared. The aim is to check the functionality and accuracy of the query.

Query 1: Date type fields: display conditions, consanguinity and inheritance pattern for patient date of birth before 2012 and Sort the results in descending order by date of birth.

Testing the query in both systems delivered the same results for both the query and the sort order. Studio 3T displays the number of the array elements, but to view the values, we have to double click on the condition elements to display it separately, as shown in Figure 9.12.



GENE2D: The Integrated Data Repository of Genetic Disorders Data\ The Query Builder : Simple Query Results

Patients: dob > 2012-01-02 : Sort by dob DESC

No	conditions	consanguinity	inheritancePattern	dob
1	Sickle cell anemia	Negative	AR	2014-07-13
2	Beta-Thalassemia	Positive	AD	2013-05-06
3	MPS IV-A	Positive	AR	2013-05-03
4	Hearing loss MPS III-B	Positive	AR	2012-11-21

The total number of patients= 4

Result | Query Code | Explain

Documents 1 to 4

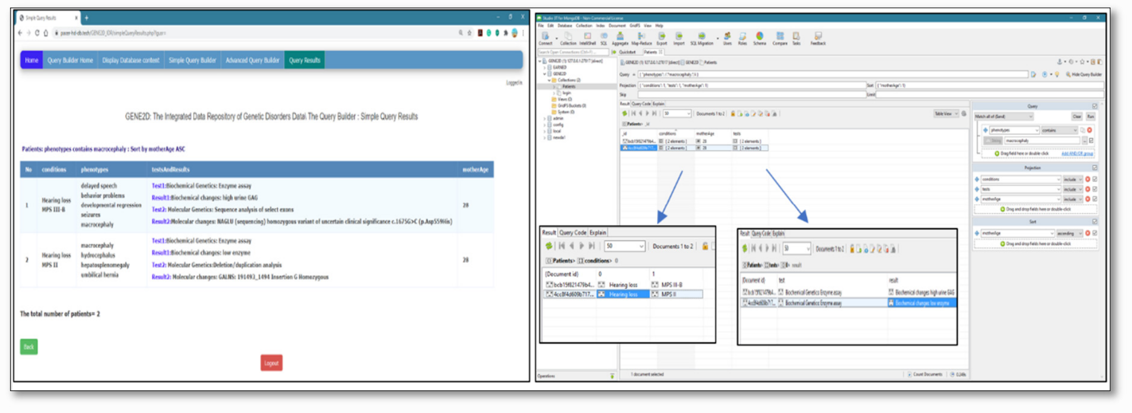
Patients> conditions> 0

(Document id)	0	1
68487dc295052a...	Sickle cell ane...	
af41e68e1309fa2...	Beta-Thalasse...	
ec4c88ca7f6953...	MPS IV-A	
bcb15f821479b4...	Hearing loss	MPS III-B

Figure 9.12: Testing query of date type field

Query 2: Array type fields: display conditions, tests and results for patients with a phenotype of "microcephaly" and sort the results in ascending order by mother age.

Results and field arrangement are the same in both systems. But as previously mentioned, Studio 3T rendering requires additional pages to display the exact contents of each array, as shown in Figure 9.13.



GENE2D: The Integrated Data Repository of Genetic Disorders Data\ The Query Builder : Simple Query Results

Patients: phenotypes contains macrocephaly : Sort by motherAge ASC

No	conditions	phenotypes	testsAndResults	motherAge
1	Hearing loss MPS III-B	delayed speech behavior problems developmental regression seizures macrocephaly	Test1:Biochemical Genetics: Enzyme assay Result1:Biochemical changes: high urine GAG Test2: Molecular Genetics: Sequence analysis of select exons Result2:Molecular changes: NAGLU (sequencing) homozygous variant of uncertain clinical significance c.1675G>C (p.Asp559His)	28
2	Hearing loss MPS II	macrocephaly hydrocephalus hepatosplenomegaly umbilical hernia	Test1:Biochemical Genetics: Enzyme assay Result1:Biochemical changes: low enzyme Test2: Molecular Genetics:Deletion/duplication analysis Result2: Molecular changes: GALNS: 191493_1494 Insertion G Homozygous	28

The total number of patients= 2

Result | Query Code | Explain

Documents 1 to 2

Patients> conditions> 0

(Document id)	0	1
bcb15f821479b4...	Hearing loss	MPS III-B
4cc8f4d609b717...	Hearing loss	MPS II

Result | Query Code | Explain

Documents 1 to 2

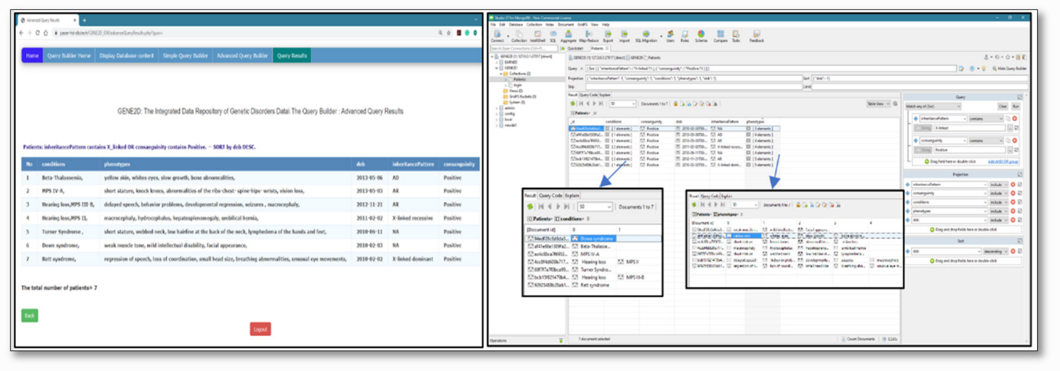
Patients> tests> 0> result

Document id	test	result
bcb15f821479b4...	Biochemical Genetics Enzyme assay	Biochemical changes high urine GAG
4cc8f4d609b717...	Biochemical Genetics Enzyme assay	Biochemical changes low enzyme

Figure 9.13: Testing query of the array type fields

Query 3: Advanced query with the logical operator (OR): display conditions and phenotypes for patients with inheritance patterns "X_linked" OR who have "positive" consanguinity; and sort the results in descending order by date of birth.

The results include all patients with "X_linked" inheritance patterns plus all patients with "positive" consanguinity. Both systems delivered the same results in the same order as presented in Figure 9.14.



GENE2D: The Integrated Data Repository of Genetic Disorders Data\ The Query Builder : Advanced Query Results

Patients: inheritancePattern contains X_linked OR consanguinity contains Positive. -- SORT by dob DESC.

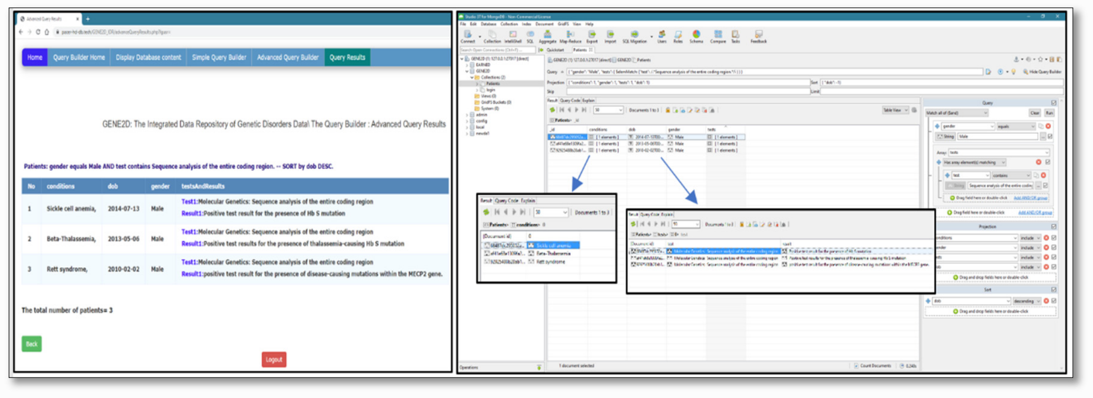
No	conditions	phenotypes	dob	inheritancePattern	consanguinity
1	Beta-Thalassemia,	yellow skin, whites eyes, slow growth, bone abnormalities,	2013-05-06	AD	Positive
2	MPS IV-A,	short stature, knock knees, abnormalities of the ribs-chest- spine-hips- wrists, vision loss,	2013-05-03	AR	Positive
3	Hearing loss,MPS III-B,	delayed speech, behavior problems, developmental regression, seizures , macrocephaly,	2012-11-21	AR	Positive
4	Hearing loss,MPS II,	macrocephaly, hydrocephalus, hepatosplenomegaly, umbilical hernia,	2011-02-02	X-linked recessive	Positive
5	Turner Syndrome ,	short stature, webbed neck, low hairline at the back of the neck, lymphedema of the hands and feet,	2010-06-11	NA	Positive
6	Down syndrome,	weak muscle tone, mild intellectual disability, facial appearance,	2010-02-03	NA	Positive
7	Rett syndrome,	regression of speech, loss of coordination, small head size, breathing abnormalities, unusual eye movements,	2010-02-02	X-linked dominant	Positive

The total number of patients= 7

Figure 9.14: Testing advanced query with the logical operator (OR)

Query 4: Advanced query field type (array of objects) with a logical operator: display conditions for male patients who have undergone gene analysis tests; and sort the result in descending order by date of birth.

The results are displayed in the GENE2D system in a table format, showing male patients who had undergone "gene analysis" sorted by date of birth. Studio 3T gives the same results but to show the array of objects (tests and their results), we have to go through multiple layers of pages, as shown in Figure 9.15.



GENE2D: The Integrated Data Repository of Genetic Disorders Data\ The Query Builder : Advanced Query Results

Patients: gender equals Male AND test contains Sequence analysis of the entire coding region. -- SORT by dob DESC.

No	conditions	dob	gender	testsAndResults
1	Sickle cell anemia,	2014-07-13	Male	Test1:Molecular Genetics: Sequence analysis of the entire coding region Result1:Positive test result for the presence of Hb S mutation
2	Beta-Thalassemia,	2013-05-06	Male	Test1:Molecular Genetics: Sequence analysis of the entire coding region Result1:Positive test results for the presence of thalassemia-causing Hb S mutation
3	Rett syndrome,	2010-02-02	Male	Test1:Molecular Genetics: Sequence analysis of the entire coding region Result1:positive test result for the presence of disease-causing mutations within the MECP2 gene.

The total number of patients= 3

Figure 9.15: Testing advanced query, field type array of objects with a logical operator

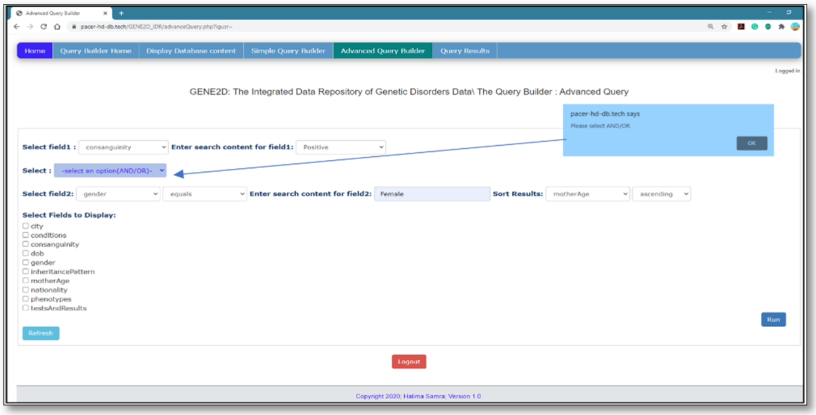
II. Interface testing

Input validation

The selection from the menus limits user errors while entering the information required for the query. The validation of data entry, especially input text boxes, is crucial, as a small entry mistake may result in a fatal error, particularly when dealing with databases such as MongoDB. Passing special characters such as (\ */| ~) may result in an unrecognised error. Therefore, we increase the data selection option and minimise data entry.

Alert messages

Warning messages: Alert messages are used to draw the user's attention to certain erroneous actions which the user has performed and informs them of what to do, for example, if the user forgets to select a field option and presses the "Run" button or on the Advanced Query Builder page if the user selects two fields and forgets to select the logical operator. An example is presented in Figure 9.16.



GENE2D: The Integrated Data Repository of Genetic Disorders Data\ The Query Builder : Advanced Query

pacer-hd-db.tech says
Please select AND/OR

Select field1 : consanguinity Enter search content for field1: Positive

Select : -select an option(AND/OR)-

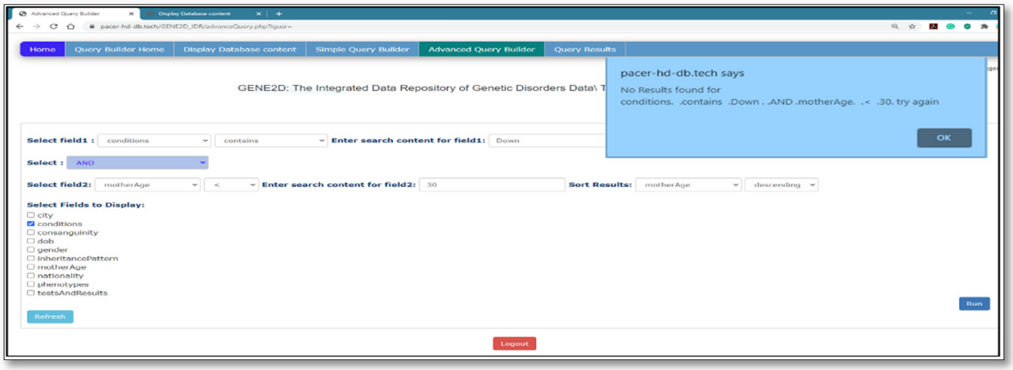
Select field2: gender equals Enter search content for field2: Female Sort Results: motherAge ascending

Select Fields to Display:
☐ city
☐ conditions
☐ consanguinity
☐ dob
☐ gender
☐ inheritancePattern
☐ motherAge
☐ nationality
☐ phenotypes
☐ testsAndResults

Copyright 2020; Halima Samra; Version 1.0

Figure 9.16: Warning messages

Informing messages: When the user submits a query request to proceed, but there are no results found for that specific question in the database, an interactive message pops up, stating the user selection, but no results are found for that selection. An example is displayed in Figure 9.17.



GENE2D: The Integrated Data Repository of Genetic Disorders Data\ T

pacer-hd-db.tech says
No Results found for
conditions. .contains .Down . .AND .motherAge. .< .30. try again

Select field1 : conditions contains Enter search content for field1: Down

Select : AND

Select field2: motherAge < Enter search content for field2: 30 Sort Results: motherAge descending

Select Fields to Display:
☐ city
☑ conditions
☐ consanguinity
☐ dob
☐ gender
☐ inheritancePattern
☐ motherAge
☐ nationality
☐ phenotypes
☐ testsAndResults

Figure 9.17: Informing messages

9.4.3.4 System deployment

I. GENE2D system deployment on a virtual private server

The first step in deploying the GENE2D system is to consider several aspects such as using multiple resources for managing the MySQL server, the MongoDB server and the web server, but all these requirements cannot be fulfilled in a shared server where all the resources are limited to what the host provides. Therefore, the system was implemented on a virtual private server to ensure system connectivity because all the resources required to run the GENE2D system are located in the same physical server.

II. GENE2D system deployment on the cloud

As the GENE2D system grows and its website traffic surges and requires more storage space than virtual private server capacity, the service must move to the cloud environment. Cloud servers use several servers in a cluster to offer unlimited storage and maximum bandwidth and to manage load balancing. So, with cloud hosting, the GENE2D system expansion will not encounter any issues due to the cloud features allocating the load to any number of machines as the system needs to balance the flow and handle the traffic, as well as provide scalability with unrestricted storage capacity, and increased reliability and performance. To this end, the GENE2D system, which is a MongoDB-based deployment on the cloud, can be managed using MongoDB Atlas with a fully automated on-demand through a pay-as-you-go model cloud service. The MongoDB Atlas server allows the deployment of the GENE2D system on cloud platforms such as Amazon Web Services, Google Cloud Platform and Microsoft Azure. Also, MongoDB Atlas offers built-in security controls and intuitive tools to work with data and extend and update the GENE2D application. It also offers visualisation tools with MongoDB Charts, reliability with distributed fault tolerance and automated data recovery, performance based on scale on-demand and elastic scalability, and efficiency through the automated deployments and database management services [212]. Based on these features, the migration of the GENE2D system from the virtual private server to the cloud, whenever needed, is a straightforward task with the MongoDB Atlas.

9.5 Results

9.5.1 NoSQL-based integrated data repository: GENE2D

The major components involved in the GENE2D architecture consist of the data sources (G3DMS), the integrated data repository as a central database, and the application interface. The GENE2D integrated data repository uses a NoSQL document store via MongoDB. The application interface called Query Builder provides multiple services for data retrieval from the database using a custom query to answer simple or complex research questions. The GENE2D integrated data repository demonstrates its potential to help grow and develop a national genetic disorders database in Saudi Arabia.

9.5.2 Design reflections

Data integration is a critical factor in the provision of large datasets for research of genetic conditions data which are collected during the diagnosis process in an individual clinic or research centre. Achieving the goal of aggregating data from multiple sources requires paying attention to different aspects in terms of design consideration and technology adoption regarding infrastructure capabilities at sources as well as an efficient storage mechanism, privacy concerns and public health support at the destination.

9.5.2.1 Integration in a low-resource setting

Integration method: The decision to select the appropriate integration method has been influenced by the resources' infrastructure capabilities. Logical integration approaches require high-performance networking, persuasive communication and infrastructure capabilities among health organisations, so the mediator software can communicate with the system that hosts the distributed data via the internet. The lack of reliable communication infrastructure in Saudi genetic clinics and research centres and the infrastructures and technologies available do not provide a suitable environment for logical integration. On the other hand, physical integration solutions such as integrated data repositories or clinical data warehouses can be constructed on top of low-resource setting systems, which do not require high-performance communication between sources, and sources being offline for some time does not affect the querying process performed on the integrated destination. Although physical integration is the best option, to be efficient and cost-effective for a low-resource environment, we have to adjust the design structure of the traditional approach of the integrated data repository which relies on standard data warehouse architecture with a predefined data model incorporated into the database schema.

Data storage model: Although the relational database model is robust storage for operational and transactional systems, as discussed in Chapter 4 Section 4.3, NoSQL databases provide more flexible options for accommodating heterogeneous data with a flexible schema model. The NoSQL document database particularly matches the requirement of the GENE2D integrated data repository to be valid for research such as flexible design according to the data access patterns, i.e. grouping data that will be read together. The decision to use a MongoDB document store as a backend for the GENE2D integrated data repository allowed us to benefit from the indexing capability, which is a crucial factor for query performance for the interface design for the GENE2D.

Extract, transform, load (ETL) process: The traditional ETL process is usually done at the staging area or transformation engine where all the complex mapping process takes place before moving data from sources to the destination. The development of ETL and customising its use is very expensive and maintenance is also a big challenge; therefore, implementing a system based on ETL is not feasible for low-resource small-and medium-sized organisations. For that reason, we decided the ETL will take place in the source of the G3DMS by adding a mapping method for extracting the required data meeting the research criteria (SQL format), transform certain fields to meet the destination specification (NoSQL format), then load to the target GENE2D. Designing and coding our own mapping method for the ETL process has several advantages: the mapping functions are designed to read data from the SQL format relational database based (MariaDB), convert and map data format to JSON format for the NoSQL document database (MongoDB) using PHP functions; the mapping method also included a de-identification method to mask the patient medical record number using a hashing function; and in addition to the indexing capability, the method also supports aggregation and a void join operation, transforming field types, and addressing multivalued fields while mapping.

9.5.2.2 Privacy compliance

The use of healthcare data for research must adhere to the Health Insurance Portability and Accountability Act privacy rule which asserts that once the data have been de-identified, the covered entities may use or disclose them without restriction, and in this case, the information is no longer considered as protected health information. Therefore, the GENE2D design managed the privacy and protected patient confidentiality in multiple design steps. First, the extraction process excluded identifiable information and documents such as patient name, photos, samples and family pedigree charts. The design included the de-identification method to be carried out at the source system under the institution’s supervision. De-identification was acquired by encoding the unique key, the patient medical record number, to solve the problem of updating the data in GENE2D while preventing duplicate entry. A hashing function was used to produce a new unique ID for each patient at the source. Finally, the GENE2D integrated data repository is only accessed by authorised users.

9.5.2.3 Supporting public health informatics

The GENE2D purpose is to create a unified view of data integrated from multiple genetic clinics and research centres. These data can be used to serve different medical research purposes. The application interface has features and methods to allow users to customise their own research queries. The system provides a different level of support from just displaying the entire datasets or using the simple query builder to investigate genetic conditions and related phenotype characteristics to support a research argument. Also, advanced query support allows combining multiple search criteria using logical operators and/or, and results can be organised and sorted efficiently. The GENE2D integrated data repository provides de-identified curated data valid to be used in genotype-phenotype research and public health-related research in Saudi Arabia.

9.5.2.4 Cost-effective implementation

Virtual private server: The first step in using the GENE2D integrated data repository is to consider several aspects such as using multiple resources for managing the MySQL server, the MongoDB server and the web server, but all these requirements cannot be fulfilled in a shared server where all the resources are limited to what the host provides. Therefore, the system was deployed on a cost-effective virtual private server to ensure system connectivity because all the resources required to run the system are located in the same physical server. The GENE2D integrated data repository was implemented on a basic virtual private server with 2 GB of RAM and a 4.8 GHz CPU, and operating system Ubuntu 18.04 64bit. These specifications provide a suitable environment to run the G3DMS and the GENE2D integrated data repository at this stage.

Cloud deployment: As the GENE2D integrated data repository grows and its website traffic surges and requires more storage space than virtual private server capacity, the service must move to the cloud environment. As the GENE2D integrated data repository is based on the MongoDB server, the deployment on the cloud can be managed using MongoDB Atlas with a fully automated on-demand pay-as-you-go model cloud service. The MongoDB Atlas server allows the deployment of the GENE2D integrated data repository on any cloud platforms such as Amazon Web Services, scale on-demand and elastic scalability, and other options. Based on these features, the migration of the GENE2D integrated data repository from the virtual private server to the cloud, whenever needed, is a straightforward task with MongoDB Atlas.

9.6 Contribution and future work

Knowing that data are available in one centralised location helps facilitate the research process for Saudi investigators in the area of genetic diseases. Our contribution is to propose a NoSQL-based integrated data repository of genetic disorders data, the GENE2D system. The purpose of the system is to integrate data from multiple local sources and provide a unified view of large datasets of genetic diagnostic data for Saudi patients with genetic disorders. Providing comprehensive datasets for data on genetic disorders in one place will assist in advancing research studies and developing and evaluating methods for diagnosis depending on patients' clinical characteristics or phenotypes, diagnostics tests and results, and family history. The GENE2D system aims to provide an easy-to-use visual query interface based on the NoSQL document database, MongoDB.

The design methodology consists of three main steps: the process of extracting, transforming and loading data. First, data is acquired from multiple G3DMSs. Then a physical schema is modelled by embedding techniques before mapping and loading the data to the GENE2D system. Next, in the implementation and testing steps, the actual system is created and tested using validation tests to examine the performance of the query and verify the system interface interaction with expected user actions. Finally, the GENE2D system is currently deployed on a virtual private server, and as the system grows, the option of cloud deployment is managed using MongoDB Atlas to accommodate any future changes and growing requirements such as increased scalability, availability, security and performance efficiency. The GENE2D system can receive data from new sources due to its flexible schema and indexing capabilities. The ultimate objective for the GENE2D system is to grow and develop the national genetic disorders database in Saudi Arabia.

9.7 Summary

This chapter was the last stage of presenting the proposed solution to the problem investigated in Chapter 2 and Chapter 3, the solution explored in Chapter 4 and Chapter 5, the solution architecture presented in Chapter 6, and the novel system G3DMS delivered in Chapter 7 and Chapter 8, with the design and implementation of GENE2D. The next chapter concludes with a summary of the work presented in the thesis and possible future works for expanding both the G3DMS and GENE2D systems.

Chapter 10: Discussion and Conclusion

10.0 Chapter overview

This chapter presents the discussion and overall analysis of the thesis findings and contributions, as well as possible future research directions emerging from this work. This chapter has the following structure. Section 10.1 introduces the chapter, Section 10.2 discusses the thesis motivation and summarises outcomes; Section 10.3 highlights some significant methods used for problem identification and definition; Section 10.4 presents the thesis outcomes; Section 10.5 asserts the thesis contribution to literature; and finally, Section 10.6 concludes the thesis and presents some future works.

10.1 Summary

The thesis focuses on the use of healthcare data for improving the care process, assisting clinical decisions, and enhancing research studies. Our research uses computer science and information concepts and methods to develop frameworks for the use of information in healthcare operations and research, as well as the creation of an integrated data repository. This field is also known as health informatics which uses health information technology to improve healthcare and uses powerful and cost-effective methods of technology to optimise the use of information in terms of the acquisition, storage and retrieval for a higher quality of services [61]. Although there is much potential for health information technology use in healthcare systems, data in databases are characterised by various and heterogeneous structures and content which presents challenges as well as opportunities [213].

Health information systems such as electronic medical records and electronic health records play a vital role in hospitals and healthcare organisations as a source for systematic data collection, storage and retrieval. Their significance emerges from the support they provide to the management and analysis of patients' data throughout the care process in a different healthcare setting. These systems do not provide an efficient source for clinical research, and their use is limited due to data quality issues as a result of inconsistent data capture methods, data integrity, completeness and accuracy [214]. In addition, other obstacles prevent the integration of data from these systems, including technical issues in database design shortcomings and biases as well as the lack of integration and sharing profile among these systems [213]. However, well-integrated health information systems are the foundation for effective and powerful clinical decision-making tools as well as reliable health information and knowledge for research [215]. Despite the evidence that successful implementation of health information systems can improve the quality of healthcare services, clinical decision making and research, their planning and implementation in developing countries face difficulties and challenges that prevent the effective use of these systems [216].

In Saudi Arabia, despite efforts by the government to encourage health organisations to accelerate the process of adopting health information systems, barriers arise at the system implementation stage [20]. These barriers include human barriers such as a lack of computer skills, staff resistance, a lack of training and consistent workflow disruption, as well as technical, financial and organisational barriers that affect adopting and implementing a health information system [217], [218], [77]. On the other hand, unsuccessful implementation is related to factors such as a lack of user satisfaction, system customizability, maintenance services and technical support [219], [80], [20]. Further challenges are encountered by health professionals who aim to use healthcare information for secondary purposes in research, such as information governance procedures, data quality, and patient privacy and legal barriers [59].

10.2 Significant research methods for problem identification and definition

The literature review on the general area of the use of health information technology in the healthcare setting in Saudi Arabia does not provide enough evidence related to the actual challenges behind the use of healthcare data for improving care process and their validity for research. Therefore, we considered further field study in a preliminary investigation of the challenges faced by researchers who conducted medical research and to what extent the health information systems within their organisations were a good source of data for their research.

Using multiple methods before addressing the actual problem helped develop a solution framework on a solid foundation of a well-identified problem. Questions distributed to healthcare professionals helped obtain the views of the actual system users and challenges related to the health information systems support to their daily workflow and when conducting research. Interviews with IT professionals who provide application support in their organisations revealed the actual health information systems capacity to support the daily care processes in addition to the ability of the system to preserve the data in an appropriate manner for use in research. The third view gained by the interview with an expert in the field added more depth to the study by providing important information about the requirements and specifications that must be available in the prospective systems to support health informatics in terms of improving services and research.

This study allowed us to narrow the field of investigation and focus on the most important application that requires more attention in Saudi society – genetic disorders, clinics and research centres that support Saudi patients from all regions including rural areas with the diagnosis and treatment of genetic diseases free of charge. A focus group interview with physicians and researchers in a genetic clinic and research centre was carried out at the first stage of the system design lifecycle, to draw more conclusions and acquire comprehensive knowledge of issues with the current systems and requirements for a new system.

10.3 Thesis outcomes

Current health information systems used in genetic research centres and clinics in Saudi Arabia have failed to enable researchers and healthcare physicians to use genetic and clinical data in their research. In addition to the scarcity of resources from which to obtain clinical and genetic data for use in research, numerous obstacles created difficulties in integrating these data from silos and scattered sources to provide standardised access to large datasets for patients with common health conditions. The research resulted in a twofold solution: a novel data management system for the diagnosis of genetic disorders called G3DMS, which is suitable for use in any genetic clinic and research centre; and a NoSQL-based integrated data repository for genetic disorders data called GENE2D to aggregate genetic conditions data from multiple genetic clinics or research centres. Research findings were published in [119], [220], [221].

Planning the design of the solution for the data management and integration systems considered several factors that may influence the implementation and adoption of the proposed systems. First, it considered the health informatics specification for the healthcare system to be able to support the workflow, assist with making decisions, and allow reuse of data collected by such systems. Second, it ensured the users of the system participated from system planning to validation and evaluation. Finally, it paid attention to infrastructure capabilities at the sources as well as an efficient storage mechanism, and privacy concerns. Table 10.1 shows the significant design features involved in the design and development of G3DMS and GENE2D.

Table 10.1: Significant design features

System	Feature	Function	Design Characteristics	Advantages
G3DMS	Informatics-enabled framework	Support the diagnosis workflow	Electronic data capture forms	Standardised data collection
			Self-developing dynamic lists	Save time, focus on the task on hand
			Systematic web pages appearance according to the diagnosis workflow for new patients	Preserve data quality and eliminate incomplete
		Support diagnosis decision	Upload and display patients' test, results, reports, published papers and photos.	Assist with confirming the provisional diagnosis cases
		Support research studies	Progressive list developed from previous user entry	Prevent data duplication a critical factor in data quality
			Customised query generation	Support patient-level or cohort investigations
			Export unidentified patient data in Excel and JSON format	Share and integrated anonymised datasets
	Significant design methods	Barker's system design method	Organise the development process in seven fundamental steps: Strategy, Analysis, Design, Build, Documentation, Transition and Production	System user involvements from the strategy phase; sequential linking between stages; smooth transition to the next stage; provide comprehensive user documentation
		Multilevel service design (MSD) method for evaluating the SDLC based on user experience	Based on MSD principles related to user experience and creating a set of interrelated	Guarantee user involvement throughout the design and implementation process; increase

System	Feature	Function	Design Characteristics	Advantages
			models that bridge the user experience and designing the service offering	user familiarity and acceptance of the system
		Informatics evaluation framework for evaluating each stage of Barker method	Based on five stages of the system development lifecycle and equivalent evaluation level according to the Stead et al. framework [208]	The evaluation process reflects knowledge to the development process; iterative design lifecycle; user satisfaction; successful implementation
GENE2D	Integration in a low-resource setting	Physical integration approach	Based on the integrated data repository (IDR) architecture	It does not require high-performance communication between sources. Sources being offline for some time does not affect the querying process performed on the IDR.
		Data storage model	Based on the NoSQL document database model, MongoDB	Inexpensive open-source software; flexible schema; indexing capabilities; high query performance
		ETL process	ETL is done at the source; self-designed extraction, mapping and loading methods	Deidentification method for masking patient MRN; convert SQL data to JSON format for NoSQL; transform complex field types; support aggregation and a void join operation

System	Feature	Function	Design Characteristics	Advantages
	Privacy compliance	De-identification method	Hashing function is used to produce a new unique ID for each patient at the source; extracting function excludes identifiable fields at the source	Adhere to the HIPAA privacy rule; provide anonymised datasets; hashing function prevents data duplication; allow multiple updates
	Support public health informatics	Query Builder interface	Design supports simple questions and complex and advanced query customisation	Provides de-identified curated datasets valid to be used in genotype-phenotype research and public health-related research in Saudi Arabia

10.4 Thesis contributions

The thesis presented major results in two main areas which addressed issues in the existing literature: data management and data integration. First, we investigated the literature regarding the use of health information systems in Saudi Arabia and the extent of their involvement in medical research. Then we examined the challenges Saudi physicians face in our survey study on the real issues hindering clinical research, such as data provision, integration and sharing among Saudi hospitals [119]. The finding highlighted the gaps in research in the defined area. We then designed the novel genetic disorders diagnosis data management system called G3DMS to be applicable in any genetic clinic or research centre [220]. Next, to solve the issue of data integration and provide large datasets for clinical research from multiple sources of genetic diagnostic data systems, we incorporated an integration framework into G3DMS using an integrated data repository based on the NoSQL database. The resulting system was the NoSQL-based integrated data repository of genetic disorders data called GENE2D [221].

This thesis attempts to bridge the research gaps identified in Chapter 4 and contribute to the health informatics initiative using our knowledge in the field of computer science and information systems to use effective concepts and methods for the design and implementation of a reliable solution to the issue of data management and integration in a particular area of the diagnosis of genetic conditions in Saudi genetic clinics and research centres.

The system is available for free use by any research centre and genetic clinic in Saudi Arabia. The main objective is to contribute to knowledge in the area of data management systems and integrated data repositories and make a positive contribution to the research and development community in Saudi Arabia and the global community.

10.4.1 Contributions to knowledge

The solutions proposed in this thesis have a significant contribution to the literature in terms of new methods, techniques, design and implementation, as follows.

I. New methods and techniques in the design and implementation of G3DMS

The contribution of G3DMS to knowledge is in its novelty as the first data management system dedicated to genetic clinics with dual functionality to serve both genetic diagnosis as well as the research process. It uses the following new methods:

- new mapping rules for uploading legacy data to the new G3DMS
- new mapping rules for extracting data from the database to Excel files

- new mapping rules for extracting data from the database to JSON format
- new Query Builder methods for managing database contents and reporting
- query-performance evaluation of G3DMS.

II. New methods in the design and implementation of GENE2D

The contribution of GENE2D to knowledge is in its pioneering use of NoSQL technology in a physical integration approach. This thesis is one of the first to use a NoSQL document-oriented database as a backend for an integrated data repository. It uses the following new methods:

- new mapping rules for extraction based on specified criteria
- new mapping rules for transforming data from an SQL source format to a Binary JSON destination format
- new mapping rules for loading and indexing data to the destination
- innovative method for the design of the integrated data repository based on the NoSQL document database (MongoDB)
- new Query Builder methods to serve both simple and complex queries in genetic-related studies and public health research
- query-performance evaluation of GENE2D.

10.4.2 Contributions to society, community and country of Saudi Arabia

I. Contributing to the Saudi genetic research community and scientific societies

Most Saudi genetic clinics and research centres are affiliated with universities and their teaching hospitals which can contribute to health research by the actual data or providing a reference to related studies in the area of genetic conditions. The proposed solution G3DMS for data management and the integration framework GENE2D will increase the contribution of these clinics and research centres to the discovery of novel genes and rare genetic diseases in the Saudi regional environment and thus contribute to international medical publications. The integrated GENE2D repository will provide a basis for research in the clinicogenomic area, such as genotype-phenotype association studies that may contribute to scientific findings or potential applications in the medical field. Furthermore, it will not only help to structure the data and facilitate the integration process but also will enable further distribution of the research data to the support community, for example, submission to public databases for clinical variation and contribute to the global community.

II. Contributing to Saudi genetic initiatives

The proposed system is designed to support recent government initiatives of the Saudi Ministry of Health in improving the national healthcare of its citizens. It also fills the gap in research and presents a reliable and practical system applicable to a low-resource setting institution with limited infrastructure capabilities of healthcare systems. The novel system for facilitating the diagnosis process G3DMS and the integrated data repository GENE2D will solve the internal institutional problems of data collection, storage and retrieval and allow data integration and sharing of healthcare data among multiple organisations with the potential to establish a national genetic disorders database.

III. Contributing to health informatics and promoting public health

There is growing awareness that secondary use of patient data for population research is the key to bridging the gap between medical research and clinical practice and moving closer to precise and personalised medicine and informed clinical decision making. The proposed system adheres to the health informatics specification to benefit from healthcare data in improving clinical workflows, assisting decision making, and reusing data in research. Therefore, the contribution will accelerate progress in human health by helping to develop a standardised data management system for effective integration of genetic and clinical data, and by stimulating data exchange rather than retention.

IV. Contributing to the research community in developing countries

Most healthcare systems in developing countries face difficulties in adopting e-health and achieve sustainable implementation of health information systems due to the lack of adequate infrastructure. Similar circumstances exist in Saudi healthcare systems, and the proposed solution considered all the aspects of designing a system applicable for a low-resource setting, as well as achieving successful and sustainable implementation. The experience of developing this system can be shared with researchers in developing countries in a collaborative effort to take advantage of the quality of healthcare, services and research in countries with similar issues.

10.5 Conclusion and future work

The efficient use of healthcare data in improving the daily care process as well as enriching research studies with efficient quality data in Saudi genetic clinics and research centres is very significant. Genetic clinics and research centres need a system with multiple functionalities to support the genetic disorders diagnosis process, physicians' decisions on diagnosing cases, and research on genotype-phenotype associations. However, current systems fail to satisfy these procedures and have resulted in a significant burden on physicians and researchers in their daily activities and have created barriers to their research studies. Several research papers discussed the implementation and adoption barriers to health information systems in Saudi hospitals and provided some recommendations for a successful application. The review of literature in the Saudi context revealed a lack of studies that provide technical solutions for the design and implementation of a patient-centred, cost-effective system that can solve current problems at small-scale or large-scale regarding a health informatics framework in terms of data collection, storage and retrieval.

This thesis aimed to fill the gaps and proposed a twofold technical solution: first, the design and implementation of a standalone new data management system for the diagnosis of genetic disorders called G3DMS which is equipped with three-dimensional functionalities to support the genetic diagnosis process and to assist in decision making to confirm the diagnosis and provide a query interface to customise and answer research questions in genetic disorders; and second, the design and implementation of a NoSQL-based integrated data repository for genetic disorders data called GENE2D which is based on a NoSQL document database for efficient integration of genetic conditions data gathered from multiple G3DMS to form a unified view of large datasets of genetic disorders data. GENE2D's primary purpose was to support genetic research and public health studies with its powerful and dynamic features to customise queries for simple and advanced research questions.

G3DMS is currently implemented in the Princess Al-Jawhara Centre of Excellence in Research of Hereditary Disorders at King Abdulaziz University Hospital. There is future linkage to the hospital information system, as the G3DMS uses the patient medical record number, which is the same identifier used in the health information systems at King Abdulaziz University Hospital. Also, there are potential implementation points at the Centre of Excellence in Genomic Medicine Research at King Abdulaziz University hospital and other independent and university-affiliated hospital centres.

Future implementation in research centres other than genetics is possible with minor changes to the data dictionary of the database and application interface to suit the requirements of the research centre.

There is future trend towards knowledge discovery based on genetic data in GENE2D to develop prediction models for genetic conditions using data mining and machine learning techniques. This can be achieved for ease of linking GENE2D to any existing data-mining suits such as MATLAB software or even NoSQL based systems like Hadoop for quantitative analysis.

The GENE2D integrated data repository is currently deployed on a private virtual server and, as the system grows, the option of cloud deployment is managed using MongoDB Atlas to accommodate any future changes and growing requirements such as increased scalability, availability, security and performance efficiency. The GENE2D integrated data repository can receive data from new sources due to its flexible schema and indexing capabilities. Therefore, many data sources of genetic disorders from any part of the country can participate in the integration framework of GENE2D. The ultimate objective for the GENE2D integrated data repository is to grow and develop the national genetic disorders database in Saudi Arabia and improve the health of the community.

References

[1] S. Finney, “Clinical Information Systems,” in *An Introduction to Clinical Governance and Patient Safety*, Oxford University Press, 2011, p. ..

[2] Institute of Medicine, Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary, Washington, DC: The National Academies Press, 2010.

[3] C. C. Yang and P. Veltri, “Intelligent healthcare informatics in big data era,” *Artificial Intelligence In Medicine,* vol. 65, no. 2, pp. 75-77, 2015.

[4] M. Khalifa, “Barriers to Health Information Systems and Electronic Medical Records Implementation A Field Study of Saudi Arabian Hospitals,” *Procedia Computer Science,* vol. 21, pp. 335-342, 2013.

[5] A. Alkraiji, T. Jackson and I. Murray, “Barriers to the Widespread Adoption of Health Data Standards: An Exploratory Qualitative Study in Tertiary Healthcare Organizations in Saudi Arabia.,” *Journal of Medical Systems,* vol. 37, no. 2, pp. 1-13, 2013.

[6] A. Alfares, “Genomics in Saudi Arabia call for data-sharing policy,” *Journal of Biochemical and Clinical Genetics,* vol. 1, no. 2, pp. 51-52, 2019.

[7] M. Al-Owain, H. Al-Zaidan and Z. Al-Hassnan, “Map of Autosomal Recessive Genetic Disorders in Saudi Arabia: Concepts and Future Directions,” *Am J Med Genet,* vol. Part A, no. 158A, pp. 2629-2640, 2012.

[8] M. Islam, T. N. Pol and Y.-C. (. Li, “Recent Advancement of Clinical Information Systems: Opportunities and Challenges,” *IMIA Yearbook of Medical Informatics,* vol. 27, no. 1, pp. 83-90, 2018.

[9] N. A. Mohamadali and N. F. A. Aziz, “The Technology Factors as Barriers for Sustainable Health Information Systems (HIS) – A Review,” *Procedia Computer Science ,* vol. 124, pp. 370-378, 2017.

[10] J. G. Teixeira, N. F. d. Pinho and L. Patrício, “Bringing service design to the development of health information systems:The case of the Portuguese national electronic health record,” *International Journal of Medical Informatics,* vol. 132, p. 103942, 2019.

[11] F. Fritz, B. Tilahun and M. Dugas, “Success criteria for electronic medical record implementations in low-resource settings: a systematic review,” *Journal of the American Medical Informatics Association,* vol. 22, pp. 479-488, 2015.

[12] R. Gliklich, N. Dreyer, M. Leavy and eds, “Registries for Evaluating Patient Outcomes: A User's Guide [Internet],” Agency for Healthcare Research and Quality (US), Rockville (MD), 2014.

[13] O. Dziadkowiec, T. Callahan, M. Ozkaynak, B. Reeder and W. J., “Using a Data Quality Framework to Clean Data Extracted from the Electronic Health Record: A Case Study,” *EGEMS (Wash DC),* vol. 4, no. 1, pp. 1-15, 2016.

[14] G. Wiederhold, Databases for Health Care, 1 ed., Berlin Heidelberg New York : Springer-Verlag , 2012.

[15] M. Sarkies, K.-A. Bowles, E. Skinner, D. Mitchell, R. Haas, M. Ho, K. Salter, K. May, D. Markham, L. O’Brien, S. Plumb and T. Haines, “Data Collection Methods in Health Services Research – hospital length of stay and discharge destination,” *Applied Clinical Informatics,* vol. 6, pp. 96- 109, 2015.

[16] M. Muji, R. Ciupa, D. Dobru, C. Bica, P. Olah, V. Bacarea and M. Marusteri, “Database Design Patterns for Healthcare Information Systems,” *MEDITECH,* vol. IFMBE Proceedings , no. 26, pp. 63-66, 2009.

[17] G. Ginsburg, “Medical genomics: Gather and use genetic data in health care,” *Nature News,* vol. 508, no. 451, p. 3, 2014.

[18] P. Nadkarni and L. Marenco, “Data Integration: An Overview,” in *Methods in Biomedical Informatics: A Pragmatic Approach*, I. N. Sarkar, Ed., Waltham, MA , Elsevier Inc, 2013, pp. 15-47.

[19] K. Alsulame, M. Khalifa and M. Househ, “E-Health status in Saudi Arabia: A review of current literature,” *Health Policy and Technology,* vol. 5, pp. 204- 210, 2016.

[20] B. Aldosari , “Rates, levels, and determinants of electronic health record system adoption: A study of hospitals in Riyadh, Saudi Arabia,” *International Journal of Medical Informatics,* vol. 83, pp. 330-342, 2014.

[21] A. El. Mahalli, “Adoption and Barriers to Adoption of Electronic Health Records by Nurses in Three Governmental Hospitals in Eastern Province, Saudi Arabia,” *Perspect Health Inf Management,* vol. 12, no. Fall, p. PMC4632875, 2015.

[22] H. Samra, A. Li, B. Soh and M. Al Zain, “Utilisation of hospital information systems for medical research in Saudi Arabia: A mixed-method exploration of the views of healthcare and IT professionals involved in hospital database management systems,” *Health Information Management Journal,* vol. 49, no. 2-3, pp. 117-126, 2020.

[23] M. Abu-Elmagd, M. Assidi, H.-J. Schulten, A. Dallol, P. N. Pushparaj, F. Ahmed, . S. Schere and M. Al-Qahtani, “Individualized medicine enabled by genomics in Saudi Arabia,” *BMC Medical Genomics,* vol. 8, no. (Suppl 1):S3, pp. 1-17, 2015.

[24] J. Kaiser, “Saudi gene hunters comb country's DNA to prevent rare diseases,” *Science,* vol. 1, no. 1, p. pp., 2017.

[25] M. Zelkowitz, A. Shaw and J. Gan, Principles of Software Engineering and Design, P.5. ed., Englewood Cliffs, NJ: Prentice-Hall , 1979.

[26] E. Coiera, F. Magrabi and V. Sintchenko, Guide to Health Informatics, London: Taylor & Francis Group, 2015.

[27] A. Alkraiji, T. Jackson and I. Murray, “The Role of Health Data Standards in Developing Countries,” *Journal of Health Informatics in Developing Countries,* vol. 6, no. 2, pp. 455-466, 2012.

[28] P. Sinha, G. Sunder, P. Bendale, M. Mantri and A. Dande, Electronic Health Record : Standards, Coding Systems, Frameworks, and Infrastructures,, Available from: ProQuest Ebook Central.: John Wiley & Sons, Incorporated, Somerset, 2012.

[29] WHO, Design and implementation of health information systems, T. Lippeveld, R. Sauerborn and C. Bodart, Eds., Geneva: World Health Organisation, 2000.

[30] WHO, “58th World Health Assembly Report,” WHO, Geneva, 2005.

[31] T. U. Zaman, T. M. Abdul Raheem, G. M. Alharbi, M. F. Shodri, A. H. Kutbi, S. M. Alotaibi and K. S. Aldaadi, “E-health and its Transformation of Healthcare Delivery System in Makkah, Saudi Arabia,” *International Journal of Medical Research & Health Sciences,* vol. 7, no. 5, pp. 76-82, 2018.

[32] K. Alsulame, M. Khalifa and M. Househ, “E-Health status in Saudi Arabia: A review of current literature,” *Health Policy and Technology,* vol. 5, no. *, pp. 204-210, 2016.

[33] J. H. Panir, “Role of ICTs in the Health Sector in Developing Countries: A Critical Review of Literature,” *Journal of Health Informatics in Developing Countries,* pp. 197- 208, 2011.

[34] J. A. Blaya, H. S. Fraser and B. Holt, “E-Health Technologies Show Promise In Developing Countries,” *Health Affairs,* vol. 29, no. 2, 2010.

[35] A. Noor, “The Utilization of E-Health in the Kingdom of Saudi Arabia,” *International Research Journal of Engineering and Technology (IRJET) ,* vol. 06, no. 09, pp. 1229- 1239, 2019.

[36] MOH, “National E- Health Strategy,” 2018. [Online]. Available: https://www.moh.gov.sa/en/Ministry/nehs/Pages/default.aspx. [Accessed 25 April 2018].

[37] B. Aldosari, “Rates, levels, and determinants of electronic health record system adoption: A study of hospitals in Riyadh, Saudi Arabia,” *International Journal of Medical Informatics,* vol. 83, pp. 330-342, 2014.

[38] A. I. Alkraiji, O. El-Hassan and F. A. Amin, “Health Informatics Opportunities and Challenges: Preliminary Study in the Cooperation Council for the Arab States of the Gulf,” *Journal of Health Informatics in Developing Countries,* vol. 8, no. 1, 2014.

[39] B. I. Blum, Clinical Information Systems, New York: Springer-Verlag. Web, 1986.

[40] A. S. Al-Mudimigh, “Successful Implementation Of Integrated Health Information Systems: The Role And Impact Of Business Process Management,” *International Journal of Computer Theory and Engineering,* vol. 1, no. 3, pp. 251- 257, 2009.

[41] A. Aljohani, P. Davis and R. Connolly, “Healthcare Information Technology (HIT) in Saudi Arabia health care systems: An overview,” in *18th annual Irish Academy of Management Conference*, Galway, Ireland, 2015.

[42] M. M. Altuwaijri, “Achieving Excellence in Electronic Health Record Deployment in Middle East Hospitals,” *Proceedings of the 4th IEEE international conference on biomedical engineering and informatics (BMEI),* vol. 4, p. 1919–23, 2011.

[43] J. Hall, J. Ryan, B. Bray, C. Brown, D. Lanfear, L. K. Newby, . M. Relling, N. Risch, D. Roden, S. Shaw, J. Tcheng, J. Tenenbaum, T. Wang and W. Weintraub, “Merging Electronic Health Record Data and Genomics for Cardiovascular Research: A Science Advisory From the American Heart Association.,” *Circulation Cardiovascular genetics,* vol. 9, no. 2, pp. 193-202, 2016.

[44] A. Baus, K. Zullig, D. Long, C. Mullet, C. Pollard, H. Taylor and J. Coben, “Developing Methods of Repurposing Electronic Health Record Data for Identification of Older Adults at Risk of Unintentional Falls,” *Perspectives in Health Information Management,* vol. 13, pp. 1-32, 2016.

[45] P. Velentgas, N. Dreyer, P. Nourjah, S. Smith and M. Torchia, “Developing a Protocol for Observational Comparative Effectiveness Research: A User’s Guide,” AHRQ Publication No. 12(13)-EHC099, Rockville, MD, 2013.

[46] J. Orechia , A. Pathak, Y. Shi, A. Nawani, A. Belozerov, C. Fontes, C. Lakhiani, C. Jawale, C. Patel, D. Quinn, D. Botvinnik, E. Mei, E. Cotter, J. Byleckie, M. Ullman-Cullere, P. Chhetri, P. Chalasani, P. Karnam, R. Beaudoin, S. Sahu, Y. Belozerova and J. Mathew, “OncDRS: An integrative clinical and genomic data platform for enabling translational research and precision medicine,” *Applied & Translational Genomics,* vol. 6, pp. 18-25, 2015.

[47] S. Alshawi, F. Missi and T. Eldabi, “Healthcare information management: the integration of patients’ data,” *Logistics Information Management,* vol. 16, no. 3/4, pp. 286-295, 2003.

[48] R. Schoenberg and S. Charles , “Internet based repository of medical records that retains patient confidentiality,” *BMJ,* vol. 321, no. 7270, pp. 1199-203, 2000.

[49] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: promise and potential,” in *Health Information Science and Systems*, http://hissjournal.biomedcentral.com/articles/10.1186/2047-2501-2-3, 2014.

[50] J.-J. Yang, J. Li, J. Mulder, Y. Wang, S. Chen, H. Wu, Q. Wang and H. Pan, “Emerging information technologies for enhanced healthcare,” *Computers in Industry,* vol. 69, pp. 3-11, 2015.

[51] M. Muji, R. Ciupa, D. Dobru, C. Bica, P. Olah, V. Bacarea and M. Marusteri, “Database Design Patterns for Healthcare Information Systems,” *MEDITECH ,* vol. IFMBE Proceedings , no. 26, pp. 63-66, 2009.

[52] O. Brazhnik, “Databases and the geometry of knowledge,” *Data & Knowledge Engineering ,* vol. 61, no. 2, pp. 207-227, 2007.

[53] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy and . P. Tarczy-Hornoch, “Data integration and genomic medicine,” *Journal of Biomedical Informatics,* vol. 40, pp. 5-16, 2007.

[54] S. Tao, L. Cui, X. Wu and G.-Q. Zhang, “Facilitating Cohort Discovery by Enhancing Ontology Exploration, Query Management and Query Sharing for Large Clinical Data Repositories,” *AMIA Annu Symp Proc,* vol. 2017, no. PMCID: PMC5977665, pp. 1685-1694, 2017.

[55] S. Murphy, “Data Warehousing for Clinical Research,” in *Encyclopedia of Database Systems*, L. LIU and M. T. ÖZSU, Eds., Boston, MA, Springer US , 2009, pp. 679- 683.

[56] V. Huser and J. Cimino, “Desiderata for Healthcare Integrated Data Repositories Based on Architectural Comparison of Three Public Repositories,” *AMIA Annual Symposium Proceedings,* vol. 2013, no. Print, pp. 648-6566, 2013.

[57] S. M. Meystre, C. Lovis, T. Bürkle, G. Tognola, A. Budrionis and C. U. Lehmann, “Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress,” *Yearb Med Inform,* vol. 26, no. 1, pp. 38-52, 2017.

[58] C. Safran, “Reuse of Clinical Data,” *Yearb Med Inform,* vol. 9, no. 1, pp. 52-4, 2014.

[59] M. H. van Velthoven, N. Mastellos, A. Majeed, J. O’Donoghue and J. Car, “Feasibility of extracting data from electronic medical records for research: an international comparative study,” *BMC Medical Informatics and Decision Making,* vol. 16, no. 90, 2016.

[60] O. H, R. C, E. M and J. A., “What Is eHealth (3): A Systematic Review of Published Definitions,” *J Med Internet Res.,* vol. 7, no. 1, 2005.

[61] F. Sullivan, “What is health informatics?,” *Journal of Health Services Research and Policy,* vol. 6, no. 4, pp. 251-254, 2001.

[62] A. Venot , A. Burgun and C. Quantin, “Medical Informatics as a Scientific Discipline,” in *Medical Informatics, e-Health: Fundamentals and Applications*, Paris, France, Springer Paris, 2014, pp. 1-10.

[63] M. Cuggia, P. Avillach and C. Daniel, “Representation of Patient Data in Health Information Systems and Electronic Health Records,” in *Medical Informatics, e-Health Fundamentals and Applications*, A. Venot , A. Burgun and C. Quantin, Eds., Paris, France, Springer-Verlag, 2014, pp. 65-89.

[64] I. N. Sarkar, “Biomedical informatics and translational medicine,” *Journal of Translational Medicine,* vol. 8, no. 22, 2010.

[65] MOH, “About the Ministry: Mission.,” 2019. [Online]. Available: https://www.moh.gov.sa/en/Ministry/About/Pages/Mission.aspx. [Accessed 10 March 2019].

[66] F. M. Albejaidi, “Healthcare System in Saudi Arabia: An Analysis of Structure, Total Quality Management and Future Challenges,” *Journal of Alternative Perspectives in the Social Sciences ,* vol. 2, no. 2, pp. 794-818, 2010.

[67] M. Almalki, G. Fitzgerald and M. Clark, “Health care system in Saudi Arabia: an overview,” *Eastern MediterraneanHealth Journa,* vol. 17, no. 10, pp. 784 -793, 2011.

[68] WHO, “Country Cooperation Strategy for WHO and Saudi Arabia 2012 - 2016,” World Health Organization, Regional Office for the Eastern Mediterranean, 2013.

[69] Ministry of Economy and Planning, “KINGDOM OF SAUDI ARABIA: MILLENNIUM DEVELOPMENT GOALS,” United Nations Development Program, New York, NY 10017 USA, 2013.

[70] M. Mufti, Healthcare Development Strategies in the Kingdom of Saudi Arabia, US: Springer, 2000.

[71] R. Hasanain, K. Vallmuur and M. Clark, “Electronic Medical Record Systems in Saudi Arabia: Knowledge and Preferences of Healthcare Professionals,” *Journal of Health Informatics in Developing Countries,* vol. 9, no. 1, 2015.

[72] E. A. Al Taisan and M. E. Seliaman , “Perceived Barriers and Drivers of Health Information Systems Adoption by Public Hospitals in Alhasa,” in *2018 21st Saudi Computer Society National Computer Conference (NCC)*, Riyadh, Saudi Arabia, 2018.

[73] QuadraMed , “Saudi Arabia Health Care System Receives Coveted “Excellence in Electronic Health Records” Award with QuadraMed’s EHR Solution,” *Business Wire,* 2010.

[74] A. K. Jabali and M. Jarrar, “Electronic Health Records Functionalities in Saudi Arabia: Obstacles and Major Challenges,” *Global Journal of Health Science,* vol. 10, no. 4, pp. 916-9736, 2018.

[75] S. H. Alkadi, “The Healthcare System in Saudi Arabia and its Challenges: The Case of Diabetes Care Pathway,” *Journal of Health Informatics in Developing Countries,* vol. 10, no. 1, 2016.

[76] H. A. Alzghaibi, “Implementing a Large-scale Electronic Health Record System in the Primary Healthcare Centres in Saudi Arabia,” *Doctoral thesis, Swansea University,* 2019.

[77] M. Khalifa, “Technical and Human Challenges of Implementing Hospital Information Systems in Saudi Arabia,” *Journal of Health Informatics in Developing Countries,* vol. 8, no. 1, pp. 12-25, 2014.

[78] R. A. Hasanain and H. Cooper, “Solutions to Overcome Technical and Social Barriers to Electronic Health Records Implementation in Saudi Public and Private Hospitals,” *Journal of Health Informatics in Developing Countries,* vol. 8, no. 1, pp. 46- 63, 2014.

[79] S. Alanazy, “Factors associated with implementation of electronic health records in Saudi Arabia,” *ProQuest Dissertations and Theses,* 2006.

[80] A. El Mahalli, “Adoption and Barriers to Adoption of Electronic Health Records by Nurses in Three Governmental Hospitals in Eastern Province, Saudi Arabia,” *Perspect Health Inf Manag.,* vol. 12, no. Fall, p. 1f, 2015.

[81] A. S. Alghamdi, “Factors associated with the implementation and adoption of electronic health records (EHRs) in Saudi Arabia,” *ProQuest Dissertations and Theses,* 2015.

[82] R. A. Hasanain, “Development of an EMR implementation framework for public hospitals in Saudi Arabia,” *Queensland University of Technology,* 2015.

[83] S. Almuayqil, A. S. Atkins and B. Sharp, “Ranking of E-Health Barriers Faced by Saudi Arabian Citizens, Healthcare Professionals and IT Specialists in Saudi Arabia,” *Health,* vol. 8, pp. 004-1013, 2016.

[84] M. M. Altuwaijri, “Supporting the Saudi e-health initiative: the Master of Health Informatics programme at KSAU-HS,” *Eastern MediterraneanHealth Journal,* vol. 16, no. 1, 2010.

[85] M. Almalki, M. Househ and M. Alhefzi, “Developing a Saudi Health Informatics Competency Framework: A Comparative Assessment,” in *MEDINFO 2019: Health and Wellbeing e-Networks for All*, L. Ohno-Machado and B. Séroussi, Eds., International Medical Informatics Association (IMIA) and IOS Press, 2019, pp. 1101 - 1105.

[86] S. Al-Ogla, “A study of hospital and medical libraries in Riyadh,Kingdom of Saudi Arabia,” *Bull Med Libr Assoc,* vol. 86, no. 1, pp. 57-62, 1998.

[87] A. Khudair and D. Bawden, “Healthcare libraries in Saudi Arabia: analysis and recommendations,” *Aslib Proceedings,* vol. 59, no. 4/5, pp. 328-341, 2007.

[88] M. Khalifa and O. Alswailem, “Hospital Information Systems (HIS) Acceptance and Satisfaction:A Case Study of a Tertiary Care Hospital,” *Procedia Computer Science,* vol. 63, pp. 198-204, 2015.

[89] SHC, “National Health Registries,” 2017. [Online]. Available: http://www.chs.gov.sa/En/HealthRecords/Pages/default.aspx . [Accessed 10 June 2017].

[90] A. Al-Zalabani, “Online sources of health statistics in Saudi Arabia,” *Saudi Med J,* vol. 32, pp. 9-14, 2011.

[91] S. Subhani and K. Al-Rubeaan, “Design and Development of a Web-Based Saudi National Diabetes Registry,” *Journal of Diabetes Science and Technology,* vol. 4, no. 6, pp. 1574-1582, 2010.

[92] B. Zaman, R. Khandekar, S. Al Shahwan, J. Song, I. Al Jadaan, L. Al Jiasim, O. Owaydha, N. Asghar, A. Hijazi and D. Edward, “Development of a Webbased Glaucoma Registry at King Khaled Eye Specialist Hospital, Saudi Arabia: A CostEffective,” *Middle East Afr J Ophthalmol,* vol. 21, no. 2, pp. 182-185, 2014.

[93] KFMRC, “King Fahd Center for Medical Research,” 2017. [Online]. Available: http://kfmrc.kau.edu.sa/Default.aspx?Site_ID=141&Lng=EN. [Accessed 11 June 2017].

[94] R. Latif, “Medical and biomedical research productivity from the Kingdom of Saudi Arabia (2008-2012),” *Journal of Family and Community Medicine,* vol. 22, no. 1, pp. 25-30, 2015.

[95] K. Alsulame, M. Khalifa and M. Househ, “E-Health status in Saudi Arabia: A review of current literature,” *Health Policy and Technology,* vol. 5, pp. 204-210, 2016.

[96] S. Almuayqil, A. S. Atkins and B. Sharp, “Knowledge Management Framework for E-Healthcare in Saudi Arabia,” *eTELEMED ,* vol. 2015, pp. 112-117, 2015.

[97] A. M. Al-Shehri, “Can Informatics Transform Public Health Practice, Research and Learning in the Kingdom of Saudi Arabia (KSA)?,” *Journal of Health Informatics in Developing Countries,* vol. 8, no. 2, 2014.

[98] F. Alharbi, A. Atkins, C. Stanier and H. A. Al-Buti, “Strategic Value of Cloud Computing in Healthcare Organisations Using the Balanced Scorecard Approach: A Case Study from a Saudi Hospital,” *Procedia Computer Science,* vol. 98, pp. 332-339, 2016.

[99] R. Kurdi, M. Aljehani, A. Subasi and S. M. Qaisar, “Cloud computing based healthcare information systems: A proposal for the Kingdom of Saudi Arabia,” in *International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 2017.

[100] A. O. Sabbagh, “A Novel Model for Managing Health Informatics in Saudi Arabia,” *Unpublished PhD Thesis. Coventry: Coventry University,* 2015.

[101] J. Zekri, “Challenges Facing Clinical Research: An Example From a Middle Eastern Country,” *Global Journal of Health Science,* vol. 10, no. 4, pp. 916-9744, 2018.

[102] H. M. Al-Dorzi, B. Khokhar, D. White and Y. M. Arabi, “Research Experience, Interest and Perceived Barriers of Clinical Staff Working at the Intensive Care Department of a Tertiary Care Academic Hospital in Saudi Arabia,” *Middle East Journal of Anesthesiology (MEJA),* vol. 22, no. 3, pp. 301- 307, 2013.

[103] M. M. Gamee, “Overview of clinical research logistics in Saudi Arabia; Barriers and difficulties facing researchers-clinical researchers’ perception,” *International Journal of Academic Research and Development,* vol. 3, no. 4, pp. 168- 175, 2018.

[104] N. Sheblaq and A. Al Najjar, “The Challenges In Conducting Research Studies In Arabic Countries,” *Open Access Journal of Clinical Trials,* vol. I, no. I, pp. 57-66, 2019.

[105] S. Al Dalbhi, A. Alodhayani, Y. Alghamdi, S. Alrasheed, A. Alshehri and N. Alotaibi, “Difficulties in conducting clinical research among healthcare practitioners in Saudi Arabia: A cross-sectional survey,” *J Family Med Prim Care,* vol. 8, no. 6, pp. 1877-1883, 2019.

[106] S. Ali, M. Alghamdi, J. Alzhrani and E. De Vol, “Magnitude and characteristics of clinical trials in the Kingdom of Saudi Arabia: A cross-sectional analysis,” *Contemp Clin Trials Commun,* vol. 3, no. 7, 2017.

[107] Y. Sato , “QUESTIONNAIRE DESIGN FOR SURVEY RESEARCH: EMPLOYING WEIGHTING METHOD,” in *ISAHP*, Honolulu, Hawaii, 2005.

[108] J. Reynaldo and A. Santos, “Cronbach’s alpha: a tool for assessing the reliability of scales,” *Journal of Extension,* vol. 37, no. 2, pp. 1-4, 1999.

[109] I. Olaronke and O. Oluwaseun, “Big Data in Healthcare: Prospects, Challenges and Resolutions,” San Francisco, 2016.

[110] H.-J. Yu, H.-S. Lai, K.-H. Chen, H.-C. Chou, J.-M. Wu, S. Dorjgochoo, A. Mendjargal, E. Altangerel, Y.-W. Tien, C.-W. Hsueh and F. Lai, “A sharable cloud-based pancreaticoduodenectomy collaborative database for physicians: Emphasis on security and clinical rule supporting,” *Computer Methods and Programs in Biomedicine,* vol. III, pp. 488-497, 2013.

[111] J.-J. Yang, J. Li, J. Mulder, Y. Wang, S. Chen, H. Wu, Q. Wang and H. Pan, “Emerging information technologies for enhanced healthcare,” *Computers in Industry,* vol. 69, p. 3–11, 2015.

[112] D. Löper, M. Klettke, I. Bruder and A. Heuer, “Integrating Healthcare-Related Information Using the Entity-Attribute-Value Storage Model,” Berlin Heidelberg, 2012.

[113] G. Iyawa, M. Herselman and A. Botha, “Digital health innovation ecosystems: From systematic literature review to conceptual framework,” *Procedia Computer Science ,* vol. 100 , p. 244 – 252, 2016 .

[114] P. Velentgas, N. Dreyer, P. Nourjah, S. Smith and M. Torchia, Developing a Protocol for Observational Comparative Effectiveness Research: A User’s Guide, Rockville: Agency for Healthcare Research and Quality, 2013.

[115] R. Richesson and J. Andrews, Clinical Research Informatics, ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/latrobe/detail.action?docID=884452. ed., London: Springer , 2012.

[116] T. Botsis, G. Hartvigsen, F. Chen and C. Weng, “Secondary Use of EHR: Data Quality Issues and Informatics Opportunities,” *Summit on Translational Bioinformatics,* vol. 2010, pp. 1-5, 2010.

[117] P. Oláh, M. Mărușteri, M. Muji, V. Bacârea, B. Haifa, M. Petrișor and D. Daniela, “A Database Design Pattern for Structuring Hierarchical Medical Data,” *Acta Medica Marisiensis,* vol. 58, no. 6, pp. 429-432, 2012.

[118] A. Alfares, “Genomics in Saudi Arabia call for data-sharing policy,” *J. Biochem. Clin. Genet.,* vol. 1, pp. 51-54, 2019.

[119] H. Samra, A. Li, B. Soh and M. Al Zain, “Utilisation of hospital information systems for medical research in Saudi Arabia: A mixed-method exploration of the views of healthcare and IT professionals involved in hospital database management systems,” *Health Inf. Manag. J.,* vol. 49, no. 2-3, pp. 117-126, 1 May 2020.

[120] F. Alkuraya, “Genetics and genomic medicine in Saudi Arabia,” *Molecular Genetics & Genomic Medicine,* vol. 2, pp. 369-378, 2014.

[121] VNR, “Sustainable Development Goals: 1st Voluntary National Review of Saudi Arabia,” United Nations Development Programme, New York, 2018.

[122] M. Abu-Elmagd, M. Assidi, H.-J. Schulten, A. Dallol, P. Pushparaj, . F. Ahmed, S. Scherer and . M. Al-Qahtani, “Individualized medicine enabled by genomics in Saudi Arabia,” *BMC Medical Genomics,* vol. 8, no. (Suppl 1), p. S3, 2015.

[123] D. A. M. A. M. e. a. Monies, “The landscape of genetic diseases in Saudi Arabia based on the first 1000 diagnostic panels and exomes,” *Human Genetics,* vol. 136, no. 8, pp. 921-939, 2017.

[124] A. Harries, R. Zachariah and D. Maher, “The power of data: using routinely collected data to improve public health programmes and patient outcomes in low- and middle-income countries,” *Tropical Medicine and International Health,* vol. 18, no. 9, p. 1154–1156, 2013.

[125] A. Boonstra, A. Versluis and J. Vos, “Implementing electronic health records in hospitals: a systematic literature review,” *BMC health services research,* vol. 14, no. 370, 2014.

[126] AMA, “Impact of High Capital Costs of Hospital EHRs on the Medical Staff,” American Medical Association, 2019.

[127] V. Palabindala, A. Pamarthy and N. Jonnalagadda, “Adoption of electronic health records and barriers,” *Journal of community hospital internal medicine perspectives,* vol. 6, no. 5, p. 32643, 2016.

[128] R. Hasanain, M. Clark and K. Vallmuur, “Progress and challenges in the implementation of Electronic Medical Records in Saudi Arabia: A systematic review,” *Health Informatics - An International Journal,* vol. 3, no. 2, pp. 1-14, 2014.

[129] I. (. o. Medicine), Clinical data as the basic staple of health learning: Creating and protecting a public good: Workshop Summary, Washington, DC: The National Academies Press., 2010.

[130] H. Lee, C. Julius, S. Ruediger, D. Rafael, W. Zhijun, G. Boris, G. Jean-Francois and L. Mingde, “How I Do IT: A Practical Database Management System to Assist Clinical Research Teams with Data Collection, Organization, and Reporting.,” *Academic Radiology,* vol. 22, no. 4, pp. 527-33, 2015.

[131] Y. K. Loke, “Use of databases for clinical research,” *Arch Dis Child,* vol. 99, p. 587–589, 2014.

[132] M. Collen, “Clinical Research Databases—A Historical Review,” *Journal of Medical Systems ,* vol. 14, no. 6, pp. 323-44. Web, 1990.

[133] T. Sahama and P. Croll , “A Data Warehouse Architecture for Clinical Data Warehousing,” in *Conferences in Research and Practice in Information Technology*, Ballarat, Victoria, 2007.

[134] S. Batra , S. Sachdeva and S. Bhalla, “Entity Attribute Value Style Modeling Approach for Archetype Based Data,” *Information,* vol. 9, no. 2, pp. 1-30, 2018.

[135] M. Collen, Computer Medical Databases: The First Six Decades (1950–2010), 1 ed., K. Hannah and M. Ball, Eds., London Dordrecht Heidelberg New York: Springer, London, 2012.

[136] P. Rob and C. Coronel, Database Systems: Design, Implementation, and Management, 8th Edition ed., Boston, Mass: Course Technology, 2009.

[137] R. Stephens and R. Plew, Database Design, - ed., Indianapolis, Indiana: Sams Publishing, 2000.

[138] G. Simsion and G. Witt, Data Modeling Essentials, Third Edition ed., San Francisco, CA: Morgan Kaufmann, 2005.

[139] M. M. Yusof, “A case study evaluation of a Critical Care Information System adoption using the socio-technical and fit approach,” *International Journal of Medical Informatics ,* vol. 84, pp. 486-499, 2015.

[140] E. deRiel, N. Puttkammer, N. Hyppolite, J. Diallo, S. Wagner, J. G. Honore´, J. G. Balan, N. Celestin, J. S. Valle` s, N. Duval, G. Thimothe´, J. Boncy, N. R. L. Coq and S. Barnhart, “Success factors for implementing and sustaining a mature electronic medical record in a low-resource setting: a case study of iSante´ in Haiti,” *Health Policy and Planning,* vol. 33, no. 2, pp. 237-246, 2018.

[141] E. Ammenwerth, C. Iller and C. Mahler, “IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study,” *BMC Medical Informatics and Decision Making,* vol. 6, no. 3, 2006.

[142] R. E. Bawack and J. R. K. Kamdjoug, “Adequacy of UTAUT in clinician adoption of health information systems in developing countries: The case of Cameroon,” *International Journal of Medical Informatics,* vol. 109, pp. 15-22, 2018.

[143] C. Moucheraud, A. Schwitters, C. Boudreaux, D. Giles, P. H. Kilmarx, N. Ntolo, Z. Bangani, M. E. St. Louis and T. J. Bossert, “Sustainability of health information systems: a three-country qualitative study in southern Africa,” *BMC Health Services Research,* vol. 17, no. 23, 2017.

[144] S. H. Afrizal, P. W. Handayani, A. N. Hidayanto, T. Eryando, M. Budiharsana and E. Martha, “Barriers and challenges to Primary Health Care Information System (PHCIS) adoption from health management perspective: A qualitative study,” *Informatics in Medicine Unlocked,* vol. 17, p. 100198, 2019.

[145] F. Kitsios, S. Stefanakakis, M. Kamariotoua and L. Dermentzogloub, “E-service Evaluation: User satisfaction measurement and implications in health sector,” *Computer Standards & Interfaces,* vol. 63, p. 16–26, 2019.

[146] I. J. Gotham, L. H. Le, D. L. Sottolano and K. J. Schmit, “An informatics framework for public health information systems: a case study on how an informatics structure for integrated information systems provides benefit in supporting a statewide response to a public health emergency,” *Inf Syst E-Bus Manage,* vol. 13, p. 713–749, 2015.

[147] J. G. Teixeira, N. F. de Pinho and L. Patrício, “Bringing service design to the development of health information systems: The case of the Portuguese national electronic health record,” *International Journal of Medical Informatics,* vol. 132, p. 103942, 2019.

[148] L. Patrı´cio, R. P. Fisk, J. F. e. Cunha and L. Constantine, “Multilevel Service Design: From Customer Value Constellation to Service Experience Blueprinting,” *Journal of Service Research,* vol. 14, no. 2, pp. 180-200, 2011.

[149] D. Kaufman, W. D. Roberts, J. Merrill, T.-Y. Lai and S. Bakken, “Applying an Evaluation Framework for Health Information System Design, Development, and Implementation,” *Nursing Research,* vol. 55, no. 2S, p. S37–S42, 2006.

[150] V. Gligorijevic´ and N. Przˇulj, “Methods for biological data integration: perspectives and challenges,” *J R Soc Interface,* vol. 12, no. 112, p. 20150571, 2015.

[151] B. Louie , P. Mork , F. Martin-Sanchez, A. Halevy and P. Tarczy-Hornoch, “Data integration and genomic medicine,” *Journal of Biomedical Informatics,* vol. 40, pp. 5-16, 2007.

[152] Z. Bicevskaa and I. Oditisa, “Towards NoSQL-based Data Warehouse Solutions,” Riga, Latvia, 2016.

[153] M. Boussahoua, O. Boussaid and F. Bentayeb, “Logical Schema for Data Warehouse on Column-Oriented NoSQL Databases,” Lyon, France, 2017.

[154] S. MacKenzie, M. Wyatt, R. Schuff, J. Tenenbaum and N. Anderson, “Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey,” *Journal of the American Medical Informatics Association : JAMIA,* vol. 19, no. e1, p. e119–e124, 2012.

[155] R. Wynden, M. Weiner, I. Sim, D. Gabriel, M. Casale, S. Carini, S. Hastings, D. Ervin, S. Tu, J. Gennari, N. Anderson, K. Mobed, P. Lakshminarayanan, M. Massary and R. Cucina, “Ontology Mapping and Data Discovery for the Translational Investigator,” *Summit on translational bioinformatics,* vol. 2010, p. 66–70, 2010.

[156] R. Yangui, A. Nabli and F. Gargouri, “Automatic Transformation of Data Warehouse Schema to NoSQL Data Base: Comparative Study,” *Procedia Computer Science,* vol. 96, pp. 255-264, 2016.

[157] D. McCreary and A. Kelly, Making Sense of NoSQL: A guide for managers and the rest of us, Shelter Island, NY: Manning Publications Co., 2014.

[158] G. Drazena and I. Coric, “NoSQL Database Phenomenon,” in *In Bridging Relational and NoSQL Databases*, Hershey, PA, IGI Global, 2018, pp. 34-93.

[159] L. Dobos, B. Pinczel, A. Kiss, G. R´acz and T. Eiler, “A comparative evaluation of nosql database systems,” *Anales Universitatis Scientiarum Budapestinensis de Rolando Eotvos Nominatae Sectio Computatorica,* vol. 42, pp. 173-198, 2014.

[160] K. Kaur and R. Rani, “Modeling and Querying Data in NoSQL Databases,” Silicon Valley, CA, USA, 2013.

[161] C. He, “Survey on NoSQL Database Technology,” *Journal of Applied Science and Engineering Innovation ,* vol. 2, no. 2, pp. 50-54, 2015.

[162] V. Abramova, J. Bernardino and P. Furtado, “EXPERIMENTAL EVALUATION OF NOSQL DATABASES,” *International Journal of Database Management Systems (IJDMS) ,* vol. 6, no. 3, pp. 1-16, 2014.

[163] K. Dehdouh, F. Bentayeb, O. Boussaid and N. Kabachi, “Using the column oriented NoSQL model for implementing big data warehouses,” Las Vegas, Nevada, USA, 2015.

[164] G. Drazena and I. Coric, “How NoSQL Databases Work,” in *In Bridging Relational and NoSQL Databases*, Hershey, PA, IGI Global, 2018, pp. 124-175.

[165] J. Runkel, “How MongoDB is Transforming Healthcare Technology,” 11 July 2017. [Online]. Available: https://www.slideshare.net/mongodb/how-mongodb-is-transforming-healthcare-technology. [Accessed 08 08 2019].

[166] P. Sadalage and M. Fowler, NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence, Upper Saddle River, NJ, USA: Addison-Wesley, 2013.

[167] P. Atzeni, F. Bugiotti, L. Cabibbo and R. Torlone, “Data modeling in the NoSQL world,” *Computer Standards & Interfaces,* 2016.

[168] G. Drazena and I. Coric, “NoSQL Data Modeling,” in *In Bridging Relational and NoSQL Databases*, Hershey, PA, IGI Global, 2018, pp. 94-123.

[169] R. Mason, “NoSQL databases and data modeling techniques for a document-oriented NoSQL database,” *Proceedings of Informing Science & IT Education Conference (InSITE),* pp. 259-268, 2015.

[170] I. MongoDB, “Data Model Design for MongoDB,” MongoDB, Inc, 2016.

[171] A. Imam, S. Basri, R. Ahmad, J. Watada and M. González-Aparicio, “Automatic schema suggestion model for NoSQL document-stores databases,” *Journal of Big Data,* vol. 5, no. 1, pp. 1-17. Web, 2018.

[172] A. Celesti , M. Fazio and M. Villari, “A Study on Join Operations in MongoDB Preserving Collections Data Models for Future Internet Applications,” *Future Internet,* vol. 11, no. 83, pp. 1-17 Web, 2019.

[173] N. Sheikh, “Chapter 7 - Analytics Adoption Roadmap,” in *Implementing Analytics*, Elsevier Inc, 2013, p. 113–127.

[174] O. Brazhnik, “Databases and the geometry of knowledge,” *Data & Knowledge Engineering,* vol. 61, no. 2, pp. 207-227, 2007.

[175] P. Ponniah, Data modeling fundamentals: a practical guide for IT professionals, - ed., Hoboken, New Jersey: John Wiley & Sons, 2007.

[176] P. Chen, “The Entity-Relationship Model-Toward a Unified View of Data,” *ACM Transactions on Database Systems,* vol. 1, no. 1, pp. 311-339, 1976.

[177] T. Boucher and A. Yalcin, Design of Industrial Information Systems, ProQuest Ebook Central. [30 July 2018] ed., San Diego: Elsevier Science & Technology, 2010.

[178] T. Teorey, S. Lightstone, T. Nadeau and H. Jagadish, Database Modeling and Design : Logical Design, Available from: ProQuest Ebook Central. [30 July 2018]. ed., San Francisco: Elsevier Science & Technology, 2011.

[179] I. A. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani and S. U. Khan, “The rise of “big data” on cloud computing: Review and open research issues,” *Information Systems,* vol. 47, pp. 98-115, 2015.

[180] M. Ercan and M. Lane, “Evaluation of NoSQL databases for EHR systems,” Auckland, New Zealand, 2014.

[181] T. Blanke, “ Digital Asset Ecosystems Rethinking crowds and cloud,” Cambridge : Elsevier Science, online resource (191 p.), 2014.

[182] V. Bhatnagar and S. Srinivasa, “Big Data Analytics,” New Delhi, 2012.

[183] R. Akerkar, “Big Data Computing,” Boca Raton, online resource (562 p.), 2013.

[184] M. Chen, S. Mao, Y. Zhang and V. C. Leung, “Big Data Related Technologies, Challenges and Future Prospects,” Springer Link, Online, 2014.

[185] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. H. Byers, “Big data: The next frontier for innovation, competition, and productivity,” McKinsey Global Institute, Online : http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation, 2011.

[186] K. Kambatla, G. Kollias , V. Kumar and A. Grama , “Trends in big data analytics,” *Journal of Parallel and Distributed Computing,* vol. 74, no. 7, p. 2561–2573, 2014.

[187] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” *AI Magazine,* vol. 17, no. 3, pp. 37-54, 1996.

[188] A. E. Hassanien, A. T. Azar, V. Snasel, J. Kacprzyk and J. H. Abawajy, “Big Data in Complex Challenges and Opportunities,” *Studies in Big Data,* vol. 9, p. 502, 2015.

[189] M. Minelli, M. Chambers and A. Dhiraj, “Big Data, Big Analytics Emerging Business Intelligence and Analytic Trends for Today's Businesses,” Wiley , online resource (216 p.), 2012.

[190] C. P. Chen and C.-Y. Zhang , “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data,” *Information Sciences,* vol. 275, p. 314–347, 2014.

[191] H. Mohanty, P. Bhuyan and D. Chenthati, “Big Data A Primer,” *Studies in Big Data,* vol. 11, p. 195, 2015.

[192] B. Richmond, Introduction to data analytics handbook, M. a. S. H. S. T. A. Center, Ed., Washington, DC: Academy for Educational Development, 2006.

[193] B. Franks, “Taming The Big Data Tidal Wave Finding Opportunities in Huge Data Streams with Advanced Analytics,” Wiley , Online resource (334 p.), 2012.

[194] J. Archenaa and E. Mary Anita, “A Survey of Big Data Analytics in Healthcare and Government,” *Procedia Computer Science,* vol. 50, pp. 408-413, 2015.

[195] R. Pethuru , R. Anupama , N. Dhivya and D. Siddhartha , “Big Data Analytics for Healthcare,” in *High-Performance Big-Data Analytics*, Switzerland, Springer International Publishing, 2015, pp. 391- 424.

[196] S. Kudyba, “Big Data, Mining, and Analytics : Components of Strategic Decision Making,” ProQuest Ebook Central, CRC Press, 2014.

[197] P. Groves, B. Kayyali, D. Knott and S. Van Kuiken, “The ‘big data’ revolution in healthcare Accelerating value and innovation,” McKinsey & Company, file:///C:/Users/a/Downloads/The_big_data_revolution_in_healthcare%20(1).pdf, 2013.

[198] C. C. Aggarwal and C. K. Reddy, “Healthcare Data Analytics,” ProQuest Ebook Central, CRC Press, 2015.

[199] M. Collen, “Secondary Medical Research Databases,” in *Computer Medical Databases The First Six Decades (1950–2010)*, K. J. Hannah and M. J. Ball, Eds., Oakland, California, Springer-Verlag London Limited, 2012, pp. 183-193.

[200] R. Campbell, “Database Design: What HIM Professionals Need to Know,” *Perspectives in Health Information Management,* vol. 1, no. 6, p. 15, 2004.

[201] D. Cardon, “Healthcare Databases: Purpose, Strengths, Weaknesses,” 2018. [Online]. Available: https://www.healthcatalyst.com/insights/healthcare-database-purposes-strengths-weaknesses. [Accessed 12 September 2018].

[202] M. D. Bhartiya S., “Exploring Interoperability Approaches and Challenges in Healthcare Data Exchange.,” in *Smart Health*, vol. 8040, Z. D. e. al., Ed., Heidelberg, Springer, Berlin, Heidelberg, 2013.

[203] T. Benson and G. Grieve, Principles of Health Interoperability: SNOMED CT, HL7 and FHIR, Third ed., London: Springer Nature , 2016.

[204] W. Khan, M. Hussain, K. Latif, M. Afzal, F. Ahmad and S. Lee, "Process Interoperability in Healthcare Systems with Dynamic Semantic Web Services," *Computing,* vol. 95, no. :, pp. 837-862, 2013.

[205] J. Anhøj, "Generic Design of Web-Based Clinical Databases," *J Med Internet Res,* vol. 5, no. 4, p. e27, 2003.

[206] PACER-HD , "Medical Services:," 2018. [Online]. Available: https://al-jawhara-center.kau.edu.sa/Pages-Medical-Services.aspx. [Accessed 17 September 2018].

[207] Apache Friends, "XAMPP Apache + MariaDB + PHP + Perl," 2019. [Online]. Available: https://www.apachefriends.org/index.html. [Accessed 25 May 2019].

[208] W. Stead, R. Haynes, S. Fuller, C. Friedman, L. Travis and B. J. R., "Designing medical informatics," *J. Am. Med. Inform. Assoc,* vol. 1, pp. 28-33, 1994.

[209] J. Klein, P. Donohoe, N. Ernst, I. Gorton, K. Pham and C. Matser, "NoSQL Data Store Technologies," Software Engineering Institute Carnegie Mellon University, Pittsburgh, PA 15213, 2014.

[210] PHP, "MongoDB driver," 2020. [Online]. Available: https://www.php.net/mongodb. [Accessed 26 January 2020].

[211] PHP, "Using the PHP Library for MongoDB (PHPLIB)," 2020. [Online]. Available: http://docs.php.net/manual/zh/mongodb.tutorial.library.php. [Accessed 27 January 2020].

[212] MongoDB, "MongoDB Atlas," 2020. [Online]. Available: https://www.mongodb.com/cloud/atlas/lp/try2?utm_source=google&utm_campaign=gs_apac_australia_search_brand_atlas_desktop&utm_term=%2Bmongodb%20%2Bcloud&utm_medium=cpc_paid_search&utm_ad=b&gclid=Cj0KCQjwzN71BRCOARIsAF8pjfi25cekdHDe_EdSq697MX4fFuulmgHDeGxGK3. [Accessed 12 April 2020].

[213] M. Bloomrosen and D. E. Detmer, "Informatics, evidence-based care, and research; implications for national policy: a report of an American Medical Informatics Association health policy conference," *J Am Med Inform Assoc,* vol. 17, no. 2, p. 115–123, 2010.

[214] P. Coorevits, M. Sundgren, G. O. Klein, A. Bahr, B. Claerhout, C. Daniel, M. Dugas, D. Dupont, A. Schmidt, P. Singleton, G. De Moor and D. Kalra, "Electronic health records: new opportunities for clinical research," *Journal of Internal Medicine,* vol. 274, pp. 547- 560, 2013.

[215] N. Hodge, "What are health information systems, and why are they important?," *Pacific Health Dialog,* vol. 18, no. 1, 2012.

[216] M. C. Azubuike and J. E. Ehiri, "Health information systems in developing countries: benefits, problems, and prospects," *Journal of the Royal Society for the Promotion of Health,* vol. 119, no. 3, p. 180–184, 1999.

[217] R. A. Hasanain, K. Vallmuur and M. Clark, "Electronic Medical Record Systems in Saudi Arabia: Knowledge and Preferences of Healthcare Professionals," *Journal of Health Informatics in Developing Countries,* vol. 9, no. 1, pp. 23-31, 2015.

[218] H. A. Shaker, M. U. Farooq and . K. O. Dhafar, "Physicians' perception about electronic medical record system in Makkah Region, Saudi Arabia," *Avicenna J Med,* vol. 5, no. 1, pp. 1-5, 2015.

[219] A. Khudair, "Electronic health records: Saudi physicians' perspective," in *Appropriate Healthcare Technologies for Developing Countries, 2008. AHT 2008. 5th IET Seminar on*, 2008.

[220] H. Samra, A. Li and B. Soh, "G3DMS: Design and Implementation of a Data Management System for the Diagnosis of Genetic Disorders," *Healthcare,* vol. 8, no. 196, 30 June 2020.

[221] H. Samra, A. Li and B. Soh, “GENE2D: A NoSQL Integrated Data Repository of Genetic Disorders Data,” *Healthcare,* vol. 8, no. 257, 4 August 2020.

Appendices

The appendices provide additional material to those included in the thesis chapters. The following table describes the appendices contents:

Chapter	Appendices	Content
Chapter 3	Appendix 3.1	The Questionnaire
	Appendix 3.2	IT specialists interview questions
Chapter 7	Appendix 7.1	Database model documentation
	Appendix 7.2	Wireframe diagrams
Chapter 8	Appendix 8.1	The SQL Script
	Appendix 8.2	G3DMS: Source code
	Appendix 8.3	The user guide/ documentation
Chapter 9	Appendix 9.1	GENE2D: Source code
Publications	Appendix P	Published papers
Note	Chapter 8 & 9 appendices please contact the author for further details.	