UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

Electronics and Computer Science

Data Quality Assessment Instrument For Electronic Health Record Systems in Saudi Arabia

By

Omar Saud Almutiry

Thesis for the degree of Doctor of Philosophy in Computer Science

February 2017

Dedicated To

My mother,

My wife,

My three children

WITHOUT THEIR SUPPORT THIS WORK WOULD NOT HAVE BEEN POSSIBLE

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

Electronics and Computer Science

Doctor of Philosophy

Data Quality Assessment Instrument For Electronic Health Record Systems in Saudi Arabia

Omar Saud Almutiry

The provision of high quality data is of considerable importance to both business and government; poor data may lead to poor decisions, so quality plays a crucial role. With the proliferation of electronic data collection by businesses and governments, there has arisen a pressing need to assure this quality. This has been recognized by both the private and public sectors, and many initiatives such as the Data Quality Initiative Framework by the Welsh government, passed in 2004, and the Data Quality Act by the United States government, passed in 2002, have been launched to improve it in those countries.

At the same time, healthcare is a domain in which the timely provision of accurate, current and complete patient data is one of the most important objectives. Instigation of a so-called Electronic Health Record (EHR), defined as a repository of patient data in digital form that is stored and exchanged securely and is accessible by different levels of authorized users, has been attracting the attention of both research and industry. EHRs allow information regarding a patient's health to be distributed among heterogeneous information systems. This evolution has added a layer of complexity in data quality, making data quality assurance a challenging issue, as the key barriers to optimal use of EHR data are the increasing quantity of data and their poor quality.

Many data quality frameworks have been developed to measure the quality of data in information systems. However, there is no consensus on a rigorously defined set of data quality dimensions. Existing dimensions are usually based on literature reviews, industrial experiences or intuitive understanding and do not take into consideration the nature of e-healthcare systems. Moreover, definitions of these dimensions vary from one data quality framework to another.

The aim of this research is to develop a data quality framework consisting of health-relevant dimensions, and data quality measures that help health organisations to enhance the quality of their data. The study provides both subjective and objective measures for assessing the quality of data.

An 11-dimensional data quality framework has been developed and confirmed by EHR stakeholders and a group of experts and data consumers. With each dimension, several associated measures have been developed to help an organisation to measure the quality of the data populating their EHR systems. Some issues linked with the measures associated with security-related dimensions have arisen during the confirmation stage. Therefore, these issues were further discussed and reviewed with security experts in order to revise the proposed framework and its measures.

Subsequently, a case study was conducted in a large hospital to examine the practicality of the proposed instrument. The instrument was used to help hospitals to assess their data. After that, the usefulness and practicality of the instrument were examined through an evaluation questionnaire distributed to quality assessment team members. Follow-up interviews with senior managers were carried out to discuss the output of the assessment and its practicality.

The contribution of this research is the development of a proper data quality framework for EHRs in the context of Saudi Arabia which resulted in 11 health-relevant data quality dimensions. An instrument was also introduced to represent all developed and confirmed measures that assess data population in EHRs.

Table of Contents

LIST OF FIGURES

List of Tables

Declaration of authorship

I, Omar Almutiry, declare that this thesis entitled 'Data Quality Assessment Instrument for Electronic Health Record Systems in Saudi Arabia' and the work presented in it are my own and have been generated by me as the result of my own original research. I confirm that:

1. This work was done wholly or mainly while a candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

- Almutiry, O., Wills, G., Crowder, R.: A dimension-oriented taxonomy of data quality problems in electronic health records. In: 13th IADIS International Conference on e-Society, pp. 98□114. IADIS, Portugal (2015) (**Awarded best paper**)

- Almutiry, O., Wills, G. & Crowder, R., 2015. A dimension-oriented taxonomy of data quality problems in electronic health records. *IADIS International Journal on WWW/Internet*, 13(2), pp.98–114. (**Extended paper**)

- Almutiry, O., Wills, G., Alwabel, A., Crowder, R., & WaIters, R. (2013). Toward a framework for data quality in cloud-based health information system. *Proceedings of 2013 International Conference on Information Society (i-Society)*(pp. 153–157).

- Almutiry, O., Wills, G. and Crowder, R. (2013) Towards a framework for data quality in electronic health records. In *Proceedings of IADIS International Conference, e-Society 2013, Lisbon, Portugal.*

Signed:
....

Date:
....

Acknowledgments

The endless thanks goes to Lord Almighty for the help and blessings he has showered on me, which has facilitated this work and enabled me to produce my research work.

I would like to express my deep gratitude to my supervisor, Gary Wills, for his guidance, sharing his invaluable experience, patience, and encouragement to complete this work, especially during hardship times. I am deeply grateful to my supervisor, Richard Crowder, for his invaluable feedback, precious guidance, and support.

I also wish to extend my gratitude to the organisations and individuals that have been involved in and cooperated with this study, answering questionnaires and actively and enthusiastically participating in this study. Special thanks go to my close friends Ayman Mualla and Mohammed Munawer for their endless support towards the completion of my PhD.

Countless thanks go to my mother, brothers and sisters for their encouragement, support and motivation to achieve my goal. I am indebted to my mother for her prayers and support during my PhD journey. I also would like to thank my father in law and mother in law for their unforgettable help and support.

I could not find the adequate words to express my deepest gratitude to my wife for her endless support, love, immense patience, sacrifice and encouragement throughout the PhD study. Grateful Thanks to my children, my joy of life, for their love.

Glossary of terms

CIHI	Canadian Institute for Health Information
DQ	Data Quality
ECM	Expectation-Confirmation Model
ECT	Expectation-Confirmation Theory
EHR	Electronic Health Record
EMR	Electronic medical record
EPR	Electronic patient record
GP	General Practitioner
HIPPA	Health Information Privacy and Portability Act (USA)
IT	Information Technology
ICT	Information and Communication Technologies
MIT	Massachusetts Institute of Technology
NEHTA	National E-Health Transition Authority (Australia)
NHS	National Health Service (United Kingdom)
NPfIT	National Health Service National Programme for IT (United Kingdom)
TDQM	Total Data Quality Management
TQM	Total Quality Management

Chapter 1 Introduction

Electronic Health Records (EHRs) are a new evolution in healthcare systems. EHRs offer benefits for patients' care as well as for governments. EHRs are a digital form of patient medical records. An EHR is defined as a repository of patient data in digital form that is stored and exchanged securely and is accessible by different levels of authorized users (Häyrinen et al., 2008). Many studies (Thakkar & Davis, 2006; Yoon-Flannery et al., 2008) have highlighted how such systems could improve the efficiency and effectiveness of healthcare and support its sustainability.

In the area of health information systems, issues and challenges have been arisen affecting widespread adoption and implementation of EHRs. Data quality assurance is a common challenge for many institutions (Botsis et al., 2010), as the key barrier to optimal use of data in EHRs is the increasing quantity of data and their poor quality. 'Fitness for use' is a widely recognised definition of data quality (Tayi & Ballou, 1998), and this takes us beyond traditional concerns with data accuracy, leading to many dimensions of data quality and thus making a multidimensional concept.

Data quality in health information systems is attracting researchers' attention. Data quality plays an important role in all applications of information systems. The private and public sectors have recognised the importance of data quality, and many initiatives such as the Data Quality Initiative Framework by the Welsh government, passed in 2004, and the Data Quality Act by the United States government, passed in 2002, have been launched to improve the quality of data in those countries (Batini et al., 2009).

The absence of a clear understanding of data quality aspects could lead to risks to the quality of healthcare services, confusion and expensive errors (Floridi, 2013). Assessing data quality is crucial, as data is a valuable asset in business strategies (Caro et al., 2008). Hence, data quality in information systems and its aspects (dimensions) have been widely discussed by many researchers (Ballou & Pazer, 1985; Tayi & Ballou, 1998; Strong et al., 1997; Wang et al., 1995; Fox et al., 1994; Levitin & Redman, 1995; Canadian Institute for Health Information, 2009; Orfanidis et al., 2004). As a result, many frameworks of dimensions to assure data quality have been introduced and developed in order to assess and enhance the quality of data. However, these frameworks have not paid sufficient attention to the dimensions needed to ensure, for example, the integrity and origin of information (its provenance). This is due to the fact that the frameworks are generic and do not reflect the nature of the domain (healthcare).

Existing frameworks of data quality dimensions are usually based on literature reviews, industrial experiences or intuitive understanding. The definition of a dimension may vary from one framework to another, as shown by the example given by Wand and Wang (1996) in their definition of accuracy. The concept of data quality depends on the actual use of the data: what may be considered high-quality data in one application may be inadequate in another (Wand & Wang, 1996). Wand and Wang (1996) also emphasize the importance of providing a design-oriented definition of data quality that reflects the nature of information systems.

Although it is possible to identify a set of dimensions common to the majority of frameworks, the definition of these dimensions differs from one framework to another. This is because the definitions attributed to dimensions are context-dependent. In the literature, the definition of dimensions is considered from different perspectives, those of database, process and user. Other frameworks assess data in an objective way, considering data as context-independent. In this case, the automation of data quality assessment is possible. Others focus on the workflow of information between data sources and end-users, while some use the suitability of information for users as the criterion.

A number of health provider organizations have implemented EHR systems in Saudi Arabia at all levels of care; primary, secondary and tertiary (Aldosari, 2014), yet only 15.8% of government-related hospitals. Among the organizations adopting EHR there is resistance to using such systems due to concerns about poor quality of data and security (Bah et al., 2011). Khalifa (2013) and Almuayqil et al. (2016) identify the barriers that hinder the adoption of such systems, showing that confidentiality and privacy are the main concerns in the adoption of EHR systems. Another issue is that the information populating such systems does not satisfy users' needs and requirements. Moreover, they point out the importance of data standarization, which causes some quality problems affecting data consumers.

1.1 Motivation for the Research

Electronic Health Record (EHR) is a promising technology believed to enhance the quality and delivery of care to patients, and support its continuity (Thakkar & Davis 2006; Yoon-Flannery et al. 2008). EHRs show a remarkable advantage over paper-based systems and locally stored EMRs in terms of access to comprehensive patients' records. They save healthcare providers time and effort, and provide an easy access to the complete medical history of a patient from one point of care by allowing information flow between providers (Heard et al. 2000). So it is unsurprising to learn that there is growing interest in the development and implementation of EHRs by healthcare bodies (Weiskopf & Weng, 2013; Botsis et al., 2010). However, the adoption of such technology encounters many challenges, ranging over legal, ethical, financial, social and technical issues.

Maintaining the high quality of data is a challenging barrier to the spread of EHRs. Many researchers perceive the problem of missing data as an obstacle that hinders the adoption of EHR amongst stakeholders (Jones & Furukawa, 2014; Cheung et al., 2013; Hacker et al., 2012). Medical error is another barrier identified in the literature. A fear of the systems having errors fed in would affect a user's attitude to information residing in EHRs (Jones & Furukawa, 2014; Cheung et al., 2013; Hacker et al., 2012; Kirkendall et al., 2013). Security and privacy are the most serious obstacles to affect the adoption of EHR amongst patients and healthcare staff. The quality of data may be determined through assessment against a set of dimensions as the quality is a multidimensional notion. Yet, the clinical research community has failed to develop a consistent taxonomy of data quality as there is an overlap of terms among existing dimensions (Weiskopf & Weng, 2013).

Therefore, the scope of this research lies in the data quality aspects from the point of EHR stakeholders in Saudi Arabia by developing a framework consisting of health-relevant dimensions and data quality measures that help health organisations to enhance the quality of their data.

1.2 Research Objective

In order to assess the data quality of EHRs properly, this research needed to determine what to measure and how to measure it. Therefore, the aim of this research was to develop a data quality framework and a data quality assessment model (instrument) to measure all aspects of quality that compromise the benefits of health data in EHR systems. To realise this, the following research questions needed to be answered:

RQ1: What data quality determinants are important for EHR stakeholders perceived data quality?

RQ2: Based on the proposed framework, what is the appropriate instrument with which to measure quality of data in EHRs?

RQ3: what are the severity factors that make data quality problems more severe?

1.3 Organisation of the Thesis

The structure of this thesis is split into nine chapters. This chapter provides a general introduction, while the rest is divided into eight further chapters. Chapter 2 provides an overview of relevant literature available regarding the concept of EHR systems, as well as their the functionalities and requirements. It also reviews data quality frameworks and their dimensions to establish an appropriate framework for EHR systems. This includes a discussion of data quality problems and the types of 'dirty data'. The 11-dimensional data quality framework and the associated data quality items are developed in the EHR context in Chapter 3. Chapter 4 describes the research methodology and research methods conducted to confirm the proposed framework. The results and findings of the empirical study are discussed in Chapter 5. An in-depth discussion about the findings and results is provided in Chapter 6. An emerging further empirical investigation of security-related dimensions is discussed in Chapter 1. Chapter 1 sketches out the methodology, findings and implications of the conducted case study. In Chapter 9, general conclusions are drawn and the direction for future research is highlighted.

Chapter 2 Literature Review

This chapter discusses EHR (Electronic Health Record) systems and the quality of their data. First, it provides an overview of EHRs, their functions and requirements. Subsequently, the chapter examines existing data quality frameworks and their dimensions. It reviews the data quality problems reviewed in the literature to develop measures for the dimensions assessed objectively and subjectively. This is to help contribute to answering the three research questions stated in Section 1.2.

2.1 Health Data and Electronic Health Record

Health data is a broad term for information relating to the health status of individuals, the health services delivered, through the aggregated data of data warehouses, knowledge-based information used for decision making, to community data utilised for policy development in the healthcare sector (Al-Shorbaji, 2001). Data, often collected at the individual patient level and known as 'unit level data', are important to support the care of the patient. Aggregated data, referred as the sum of 'unit level data', are considered good sources for healthcare organisations and governments, and contribute to the good of society. Data play important roles during diagnose processes. High quality data and their meaningful interpretation are essential to healthcare providers to support the proper care of patients. Data in healthcare have many different formats, and are used in many ways. Data, for instance, may be textual, numerical measurement, narrative, and both still pictures and videos.

Health data can fall into two major types; clinical data or administrative data. Clinical data concerns patient-related data, captured for the purpose of healthcare provision for subjects of care. It includes the characteristics of patients, illness and the process of healthcare, while administrative data refers to information on cost and services needed for delivering care. These are usually collected by government departments or other organisations for the purpose of registration, transactions and record keeping.

Many technologies have emerged to capture, manage or transmit information on the health of individuals or the activities of organisations that work within the health sector. Health information systems have existed for about two decades, and are exemplified by hospital information systems. Their main purpose is to contribute to high-quality, efficient patient care. E-health is an emerging healthcare delivery model shaped by the evolution of Information and Communication and Technology. It denotes health services and health informatics delivered through Internet and supporting technologies. It covers a wide range of systems and services that increase efficiency in health delivery. An EHR system is considered an important resource amongst these introduced in e-health.

Many terms define patient-related electronic information in e-health services. The terms Electronic Health Record (EHR), Electronic Medical Record (EMR) and Patient Health Record (PHR) are often used interchangeably in the healthcare field, despite vital differences between them. The following is a brief discussion of electronic health record systems (EHR) and similar terms, their benefits, EHR content and functions, their requirements and some barriers to EHRs.

2.1.1 Electronic Health Records

An electronic health record, or EHR, refers to a repository of patient data records in digital form. This record can be stored and exchanged securely and is accessible by different levels of authorized users at the point of care (Häyrinen et al., 2008). Figure 2.1 shows the different inputs contributed into EHRs. Many types of medical records, such as "primary care settings" could be good inputs for EHR. Primary care is the first-contact, continuous and day-to-day healthcare given by healthcare providers (Starfield, 1994).

Another definition of EHR given by Wager et al. (2009) follows, and differentiates EHR from EMR:

> *An electronic record of health-related information on an individual that conforms to nationally recognized interoperability standards and that can be created, managed, and consulted by authorized clinicians and staff across more than one healthcare organization.*

Although EHRs are defined and discussed above, we need to clarify some EHR-related terms to achieve full understanding. ISO/TR 20514:2005 gives us a formal definition of the scope and purpose of an integrated care EHR (ISO 2005):

> *repository of information regarding the health status of a subject of care, in computer processable form, stored and transmitted securely and accessible by multiple authorized users, having a standardized or commonly agreed logical information model that*

> *is independent of EHR systems and whose primary purpose is the support of continuing, efficient and quality integrated health care*

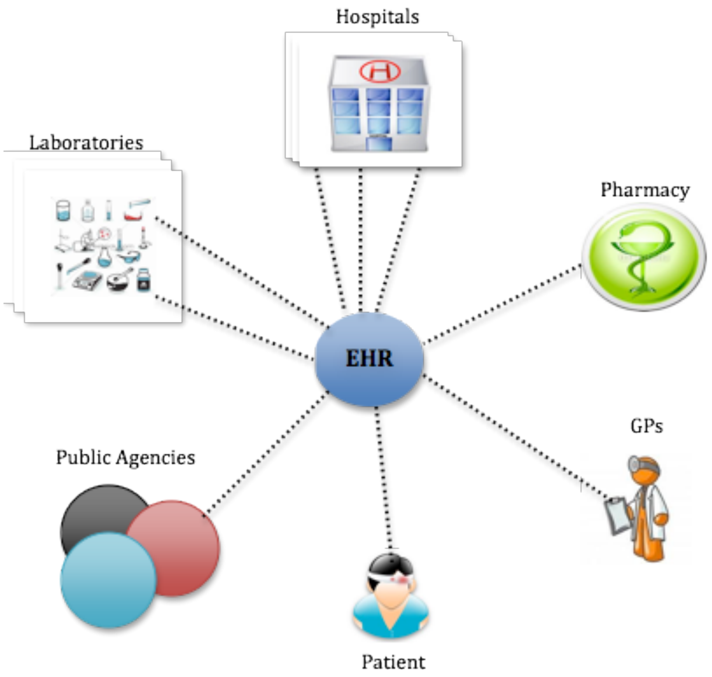


Figure 2.1. Inputs to EHR include GPs, patients, public agencies (including social care), laboratories, hospitals and pharmacies

Another term used in the EHR context is EHR information, which implies all patient health-related information that could reside in different systems. This information constitutes the patient's EHR (van der Linden et al., 2009). The rapid adoption of EHRs is promoted and influenced by many factors (Wimalasiri et al., 2004), for example:

- Several healthcare providers for patients depending on their particular healthcare needs (GPs, specialists, and physiotherapists).
- Medical records of a patient are distributed due to services being provided at disparate locations.
- Patients become more involved in their care decisions.
- Patients tend to be more mobile.

Some people conflate EMR and EHR, in spite of the fact that they describe completely different concepts (Garets & Davis, 2006). The following is a brief explanation about the two similar terms to EHR.

2.1.1.1 Electronic Medical Records

Electronic Medical Records, or MRs (Wager et al., 2009; Garets & Davis, 2006), comprise a type of application environment of electronic records of health-related information, such as clinical data, order entries and pharmacy information. Health stakeholders use these databases to document, monitor and manage care delivery within a Care Delivery Organisation (CDO). The data in an EMR is a legal record owned by the CDO and audits what happens to patients during their encounters in the healthcare organisation. EMRs are widely used in North America and Japan but are regarded as outdated by many (Kim & Lehmann, 2009).

2.1.1.2 Personal Health Records

A Personal Health Record (PHR) is defined by some researchers (National Alliance for Health Information Technology, 2008; Wager et al., 2009) as an electronic record of an individual's health-related information drawn from heterogeneous sources and managed and controlled by the individual. Such a record must comply with nationally recognized interoperability standards.

2.1.2 Benefits of EHR implementation

Information and Communication Technology (ICT) has shaped the healthcare field, leading to rapid changes in Health Information Technology (HIT). Nowadays, EHR systems are core elements of e-health systems, which, in turn, are the basis of health services around the world (Blobel & Pharow, 2009; Jahanbakhsh et al., 2011).

An enhancement of the quality of care is a noticeable advantage of adopting EHR systems, as it has been argued that EHRs tend to provide comprehensive patient records, compared to EMR or paper-based systems. Many studies (Thakkar & Davis, 2006; Yoon-Flannery et al., 2008) have highlighted how such systems could enhance quality of care and support its continuity, since access to the whole patient's record facilitates the making of more accurate and compete medical diagnoses. Moreover, EHR use promotes patient safety by reducing medical errors in hospitals (Bates, 2000; Bates et al., 1998). Medical errors can lead to deaths, of which there are an estimated 98,000 each year in the United States, costing as much as $29 billion (Hoffman & Podgurski, 2008; Institute of Medicine, 2000). EHR systems could also notify patients about important changes in drug therapy (Jain et al., 2005).

Furthermore, EHRs save health delivery organisations time and effort, and provide an easy access to the complete medical history of a patient from one point of care (Heard et al., 2000). EHRs also make the healthcare processes efficient since they allow information flow between providers. It is worth mentioning that patients' access to their EHRs is one of the significant characteristics of EHR as it allows and promotes communication between healthcare providers and patients. This experience of access leads to a better relationship with doctors, improved self-care and improved understanding of health information (Zheng & Yu, 2016; Esch et al., 2016).

Overall, EHRs show a remarkable advantage over paper-based systems and locally stored EMRs in terms of access to comprehensive patients' records. EHRs facilitate required access by healthcare professionals in various locations who are involved in the patient's care and need to cooperate and exchange data among themselves.

2.1.3 EHR challenges

The adoption of EHRs encounters many challenges, ranging over legal, ethical, financial, social and technical issues. Data interoperability is considered the major technical barrier to the adoption of EHRs, as patients' data is managed by many healthcare providers and kept in heterogeneous formats such as relational database and unstructured document storage. To overcome this issue that results in an interoperability problem, many international standards have been developed to ease exchange, integration, sharing and retrieval of electronic health information, such as HL7 Clinical Document Architecture (CDA), openEHR and CEN EN 13606 EHRcom (Dolin et al., 2006; Kalra, 2006; Begoyan, 2007; Dantu et al., 2007). Besides, as some healthcare providers may lack of secure network infrastructure, secure data communication cannot always be accomplished (Dantu et al., 2007).

Maintaining the high quality of data is a challenging barrier to the spread of EHRs. Many researchers perceive the problem of missing data as an obstacle that hinders the adoption of EHR amongst stakeholders (Jones & Furukawa, 2014; Cheung et al., 2013; Hacker et al., 2012). Medical error is another barrier identified in the literature. A fear of the systems having errors fed in would affect a user's attitude to information residing in EHRs (Jones & Furukawa, 2014; Cheung et al., 2013; Hacker et al., 2012; Kirkendall et al., 2013). Security and privacy are the most serious obstacles to affect the adoption of EHR amongst patients and healthcare staff. Failure to satisfy security and privacy concerns over EHRs would shake the stakeholders of the EHR system and may lead to patients failing to disclose important information that is required for diagnosis and treatment (Meier, 2002).

2.1.4 National EHR initiatives

Many countries have launched initiatives to use data and technologies better to contribute to improving their healthcare services by introducing EHRs. The benefits potentially to patients and public health attracted governments' investments in these projects. However, the success of EHR implementation does not rest only on addressing technical issues, as non-technical factors also need to be considered.

In Australia, the government has introduced a non-profit organisation, named the National E-Health Transition Authority NEHTA, responsible for identifying and developing the necessary foundations and technologies to deliver the best e-health system. The federal government has also announced the establishment of the Australian Digital Health Agency[1] to ensure that the nation's health system is technologically up to date. NEHTA, established in 2005, has introduced an Individual Electronic Health Record (IEHR) to develop a secure record of an individual's medical history. My eHealth Record (MyEHR) has been also introduced to allow the secure sharing of an individual's health information among their healthcare providers (National E-Health Transition Authority, 2015). NEHTA stresses the importance of the data quality principle. It urges healthcare providers to ensure that their submitted data need to be free of errors, up to date and safe for use by the medical practitioners. Data sharing is a clear national objective of the Australian government.

Canada Heath Infoway[2] is an independent, non-profit organisation, established in 2001 and funded by the federal government. It is tasked to ensure the development of interoperable EHR systems, and supporting the development of the provincial EHR infostructures (EHRi). EHRi is set to enable communication facilitation between different healthcare providers, promoting access to patient-centric clinical information (Canada Health Infoway, 2009). Infoway produced a national framework called EHR Blueprint to guide the development of EHR systems in each province. EHR Blueprint consists of guides, principles and components necessary for the interoperable EHR (Canada Health Infoway, 2006).

UK's National Health Service National Programme for IT (NPfIT) is one of the world's largest information civil technology initiatives to improve the quality of healthcare in England, costing 12.4 billion over ten years (Brennan 2005). It was initiated in 2005 to upgrade the National Health Service (NHS) to centrally mandated electronic health records for patients, and to connect thousands of general practitioners to hundreds of hospitals, providing them with secure access patient data. NPfIT was divided into a number of deliverables: a national healthcare network; an electronic appointment booking system; a large national

[1] https://www.digitalhealth.gov.au/

[2] https://www.infoway-inforoute.ca/en/

[3] http://www.connectingforhealth.nhs.uk/

[2] https://www.infoway-inforoute.ca/en/

[4] https://www.igt.hscic.gov.uk/

healthcare data repository; and five local service providers (LSP), covering the five areas (clusters) of England. The programme defined a set of standards and frameworks called NHS Interoperability Toolkit (ITK) to ease the interoperability issues and to allow data sharing between the local systems.

2.1.5 EHR Architecture Requirements

EHR requirements are intended to define generic features to make EHR communicable and complete and to retain integrity across systems. The British Standards Institute (2011) principally publishes requirements for EHR and classifies them into four themes: requirements for representation of clinical information; communication and interoperability requirements; ethical and legal requirements; and fair information principles. These must be met by the architecture of a system that processes EHR information. The following briefly discusses each type of requirement mentioned in ISO 18308.

Requirements for representation of clinical information

These requirements specify EHR features required to support the process of clinical care and its documentation. They specify the types and structure of health record entries. They also define the representation of content and data values within health record entries. Furthermore, they set guidelines for data retrieval to comply with EHR architecture. Finally, representation and support of clinical processes and workflow are discussed.

Communication and interoperability requirements

These requirements emphasize the importance of supporting retrieval of data in its original language, clearly determining authorship and time and place of creation. They also express the need to keep an audit trail of EHR communication. Furthermore, they state that EHR must support interoperability standards.

Ethical and legal requirements

The following list provides an overview of ethical and legal requirements:

- Health record provenance
- Unambiguous identification of subject of care
- Authorship of health record entries
- Identifying health record locations
- Dates and times of health care, and of the recording entries
- Version management.

Fair information principles

These requirements highlight some important principles needed to safeguard and assure fair use of EHR information. The following list includes the fair principles requirements:

- Accountability
- Identifying purposed of uses and collection
- Consent
- Access policies
- Auditability
- Subject access.

2.1.6 EHR Content and Functionality

The Institute of Medicine (IOM) Committee has identified key components of EHR systems and highlighted EHR functions (2003). These core functions fall into eight categories and are discussed briefly below.

Health information and data

Health information means health data that 'have been organised into a meaningful format, that is in such a way that [these data] can be understood and retrieved when needed', irrespective of the aggregation level (Davis & LaCour, 2014; Cabitza & Batini, 2016). EHR systems should hold a defined data set that includes, for example, medical and nursing diagnoses, allergies, demographics and laboratory test results, to ensure improved access for care stakeholders to needed information.

Results management

This feature manages results of all types, such as laboratory test results and radiology procedure results reports. This prevents redundant and additional testing, thus improving efficiency of treatment and decreasing cost.

Order entry/order management

Computerised provider order entry (CPOE) for areas such as electronic prescribing can improve workflow processes, prevent lost orders and eliminate ambiguities caused by illegible handwriting.

Decision support

Computerised decision-support systems have demonstrated the ability to enhance clinical performance in many aspects of health care through, for instance, drug alerts, rule-based alerts and reminders.

Electronic communication and connectivity

Effective communication is crucial to providing high-quality healthcare. Communication may be among healthcare team members, patients and other partners, such as pharmacy, laboratory and radiology. This communication and connectivity include the medical record integrated within the same facility, among different facilities within the same healthcare system and between different systems (Thakkar & Davis, 2006).

Patient support

Many forms of patient support have shown significant effectiveness in healthcare in general. These forms include patient and family education and home tele-monitoring.

Administrative processes

Electronic scheduling systems for hospital admission, inpatient and outpatient procedures and visits play an important role not only in enhancing the efficiency of healthcare units but in providing better services to patients.

Reporting and population health management

This feature makes the process of reporting less labour-intensive and time-consuming. It helps to report patient safety and quality data, and public health data.

2.1.7 EHR implementation in some countries

Governments have priority areas for EHR implementations and anticipated benefits. Some countries have focused on primary care like Spain, while others like England has concentrated on secondary care. The national approaches to achieve the exchange of healthcare information also vary as some countries advocated systems standardisations whereas others adopted and used interoperability standards to ease the integration concept between the existing and new IT systems (Morrison et al., 2011).

Coiera (2009) introduces a theoretical lens in which the national approaches of EHR implementation are ‘top-down’, ‘bottom-up’ or ‘middle-out’. A ‘top-down’ model is led by government with aim of central database and shared EHRs. It promotes the idea of central procurement of standardised health (IT) systems and replaces diverse existing systems. Whereas ‘bottom-up’ approach relies on the effort of local healthcare providers whom in charge of making their own decisions regarding their healthcare IT systems and their compliance with interoperability standards. England and Unite State are examples of the two mentioned strategies respectively. The last approach, ‘middle-out’, combines elements from the other two models. It combines local consultation with regard to systems choices and investment with government support with regard to nationally agreed interoperability standards (Morrison et al., 2011; Coiera, 2009).

In England, National Health Service (NHS) initially planned to deliver standardised EHR applications. This implies that the national strategy followed in England was top-down as the government was in charge of delivering the IT solutions through standardised EHR and NHS Trusts (local NHS organisations) were to stick to this national programme instead of purchasing or developing their own IT solutions for EHRs.

In United States, the assurance of EHR interoperability relies on the implementation and use of systems locally chosen. The role of federal government falls on setting government policy objectives and strategies regarding privacy and security, interoperability, adoption and data collaborative governance. Multiple users and stakeholders including patients are encouraged actively participate in the policy development process at local, state and federal levels (Morrison et al., 2011).

2.1.7.1 Rates and levels of hospital EHR system adoption

The adoption rates of EHR in hospitals are lower than those in in primary care settings (Jha et al., 2008). With regard to national EHR systems, most European countries are still at planning stage. Denmark, Sweden, the Netherlands and Switzerland are considered as European leaders in in nationwide EHR development, while Australia’s nationwide EHR system is considered one of the most advanced systems worldwide (Fragidis & Chatzoglou, 2017; Osborn & Schoen, 2013).

Table 2-1 shows the nationwide EHR implementation status of some countries (Fragidis & Chatzoglou, 2017).

Table 2-1: EHR implementation status of some countries

Country	EHR Coverage	Next Milestone
Germany	EHRs are implemented in hospitals. However, the integration between hospitals are not yet implement due to incompatibility between various systems used by these hospitals.	A new medical chip card system will be fully operable by 1 July 2018.
Netherlands	Almost all primary care settings and pharmacists are connected to National Switch Point, but not all hospitals.	All hospitals to be connected to National Switch Point.
United Kingdom	Summary Care Record system (current medication, adverse reactions, and allergies) is kept for all patients.	Universal digital care records to be implemented by 2018 for some services (urgent and emergency care services) and by 2020 for the rest of services.
Australia	2.5 million patients use the Person Controlled Electronic Health Record (PCEHR) and 8000 healthcare providers have registered.	A new commission introduced to be responsible for e-health implementation. Practice Incentives Program (PIP) was introduced to encourage GPs to participate.
Canada	EHR Interoperability across country is limited as access to EHR systems varies depending territory authority.	Establishing pan-Canadian common standards and interoperability.
New Zealand	Level of Interoperability among EHR systems is low.	A national patient portal to be implemented.
United states	Healthcare providers face low interoperability. Also, the interoperability level between healthcare providers and insurance is still low.	The Meaningful Use Incentive Program is launched to raise EHR functionality and improve data exchange.

2.2 Data Quality

Data quality is an increasingly important requirement in information era, and is regarded as the heart of an information system (Strong et al., 1997; Orr, 1998). Its importance is increasingly attributable to the advanced development of Information and Communication Technology (ICT), as well as information systems. Organisations' reliance on data exchange promotes the need for data quality assurance, as poor quality data negatively impact their businesses (Paulson, 2000; English, 1999).

2.2.1 Data quality definitions and implications

Despite its importance, there is no agreement on the definition of data quality, as it has been characterised as a multidimensional concept (Klein & Rossin, 1999; Batini & Scannapieco, 2006; Redman, 1996; Wang & Strong, 1996; Wand & Wang, 1996). It can be defined from different perspectives such as data consumers, data producers, and data custodians. 'Fitness for use' is a widely recognised definition of the quality of data (Tayi & Ballou, 1998), as this is a consumer-focused view, and takes us beyond traditional concerns with data accuracy and with the many dimensions of data quality. It recognises data quality as the ability of data to meet consumers' requirements (Wang, 1998; Orr, 1998). Data quality includes not only data validation and verification, but also the appropriateness of use (Orfanidis et al., 2004). This definition emphasises the importance of consumers' perspectives. Data quality is contextual, meaning that consumers define what would be good data quality for data use, within it context of use (Strong et al., 1997). Thus, Redman (2001) defines the data quality as:

> *Data are of high quality if they are fit for their intended uses in operations, decision-making, and planning. Data are fit for use if they are free of defects and possess desired features.*

Data quality is also defined as the degree of agreement between the data views populating an information system and that same data in the real world (Orr, 1998). Data quality of 100 per cent would imply perfect agreement, however the real concern with data quality is sufficient to ensure that data is accurate, consistent and concurrent enough for organisations to make reasonable decisions since no serious information system would achieve 100 per cent data quality. Data quality enhancements cannot be achieved only through the information technology, as the data quality problem might be caused by processes, people or technology. For example, a survey conducted by Data Warehouse Institute and surveying 647 data warehousing professionals shows that 76% of data quality issues were caused by stakeholders incorrect data entry (Eckerson, 2002).

2.2.2 Data quality context

Context is the circumstances and settings that define the type of data collected and how that data would be used, such as financial data and monitoring data (Dravis, 2004). Data context is crucial in data quality improvements, as it eases the understanding of the data quality problems that an organisation encounters. Data quality depends on the context, as data consumers evaluate the quality of the data depending on the tasks at hand (Strong et al., 1997) and the same data may be required for different tasks with varying quality characteristics. So, the quality characteristics will change over time as the organisational context may change from time to time, and any change in work requirements (Lee, 2003; Strong et al., 1997).

2.2.3 Information/data stakeholders

Data quality assessment can be influenced by information production roles along with task and context (Even & Shankaranarayanan 2007). Information production roles are recognised by Lee et al. (2002) as:

- Data custodians (known as IS professionals): those in charge of data storage and maintenance through information system design, development and maintenance.
- Data collectors: those in charge of data, supplying information production.
- Data consumers: those who use and consume data or information for a particular task.

However, a single individual may has more than a single role within an organisation, such as a nurse collecting and consuming data in a hospital. There might be differences between roles in terms of their perception towards the quality of the same data (Strong et al., 1997). Information production roles, having knowledge about the context of data, could lead a better data quality, as research (Lee & Strong 2003; Kmietowicz 2004) has shown that data collectors with domain knowledge and knowledge about why data are collected contribute to improved data quality.

2.2.4 Data quality dimensions

The definition of quality of data states that it is a multi-dimensional concept. The definition of quality widely adopted is focused on the consumer's needs and the product's fitness for use (Juran, 1988). This implies the importance of an understanding and consideration from a customer's perspective during the data quality process. Thus, the researchers have identified a number of data quality dimensions. A dimension refers to a characteristic that captures a specific facet of quality. Dimensions are defined as a set of quality attributes that represent a single construct (Wang & Strong, 1996). Dimensions are practical constructs of data quality that need to be defined and measured. Identifying and defining dimensions and respective measures are crucial activity to assess data. Generally, several measures can make up a dimension but, in some cases, a unique measure may be associated with one dimension.

In the literature, there are different classifications of quality dimensions with a number of discrepancies in the definitions of most (Batini et al., 2009). Table 2-2 summarises some proposed frameworks of data quality dimensions for information systems in general, along with their sources.

One of the most important classifications is provided by Wang and Strong (1996), who classify the identified dimensions into four categories: intrinsic, accessibility, contextual and representational, capturing the aspects of data quality that are important to data consumers. First, they gathered 179 data quality attributes from literature, researchers and data consumers. After the second survey, the number of attributes shortened to 15 data quality dimensions, as displayed in Table 2.2 along with brief description for each dimension.

Table 2-2: Data quality dimensions in information systems

Source	Classification	Data Quality Dimensions
Wang & Strong 1996	Intrinsic IQ	Accuracy, objectivity, believability, reputation
	Contextual IQ	Relevancy, value-added, timeliness, completeness, appropriate amount
	Representational IQ	Understandability, interpretability, concise representation, consistent representation
	Accessibility IQ	Accessibility, ease of operation, security
Zmud 1978	Intrinsic IQ	Accurate, factual
	Contextual IQ	Quantity, reliable
	Representational IQ	Arrangement, readable, reasonable
	Accessibility IQ	
Jarke & Vassiliou 1997	Intrinsic IQ	Believability, Accuracy, Consistency, Completeness
	Contextual IQ	Relevance, Non-volatility, Timeliness, Usage, Source currency, Data warehouse currency
	Representational IQ	Interpretability, Version control, Origin, Syntax, Semantics, Aliases
	Accessibility IQ	Transaction, Availability, Privileges, Accessibility, System Availability
DeLone & McLean 1992	Intrinsic IQ	Accuracy, Reliability, Freedom from Bias, Precision
	Contextual IQ	Completeness, Content, Currency, Relevance, Importance, Usefulness, Timeliness, Informativeness, Sufficiency
	Representational IQ	Understandability, Appearance, Clarity, Conciseness, Comparability, Format, Readability, Uniqueness
	Accessibility IQ	Quantitativeness, Convenience of Access, Usableness
Ballou & Pazer 1985	Intrinsic IQ	Accuracy, consistency
	Contextual IQ	Completeness, timeliness
	Representational IQ	
	Accessibility IQ	
Fox et al. 1994	Intrinsic IQ	Accuracy, consistency
	Contextual IQ	Completeness, currentness
	Representational IQ	
	Accessibility IQ	
Wang et al. 1995	Intrinsic IQ	Believability
	Contextual IQ	Usefulness
	Representational IQ	Interpretability
	Accessibility IQ	Accessibility

Lee et al. (2002) summarised academic and practitioners' research on data quality aspects within organisations. These dimensions were grouped into four categories introduced by (Lee et al., 2002), that is, intrinsic, accessibility, contextual and representational. Intrinsic data quality denotes that data have quality in their own right, while contextual data quality highlights the requirement that this needs to be taken into consideration within the context of the task at hand. Accessibility and representational stresses the importance of computer systems that present data and information in a way that it is interpretable, understandable and easy to manipulate, and that the system must be accessible and secure.

Table 2-3: Data quality dimensions from (Wang & Strong 1996)

Data quality dimensions	Description	category
Believability	The extent to which date are accepted or regarded as true, real, and credible.	Intrinsic data quality
Accuracy	The extent to which date are correct, reliable, and certified free of error.	
Objectivity	The extent to which date are unbiased (unprejudiced) and impartial.	
Reputation	The extent to which date are trusted or highly regarded in terms of their source or content.	
Value-Added	The extent to which date are beneficial and provide advantages from their use.	Contextual data quality
Relevancy	The extent to which date are applicable and helpful for the task at hand.	
Timeliness	The extent to which the age of the data is appropriate for the task at hand.	
Completeness	The extent to which date are of sufficient breadth, depth, and scope for the task at hand.	
Appropriate Amount of Data	The extent to which the quantity or volume of available data is appropriate.	
Interpretability	The extent to which date are in appropriate language and units, and the data definitions must be clear.	Representational data quality
Ease of Understanding	The extent to which date are clear, without ambiguity, and easily comprehended.	
Representational Consistency	The extent to which date are always be presented in the same format and compatible with previous data.	
Concise Representation	The extent to which date are compactly represented without being overwhelming.	
Access Security	The extent to which access to data can be restricted, and hence, kept secure.	Accessibility data quality
Accessibility	The extent to which date are available or easily and quickly retrievable.	

Table 2.3 lists the intrinsic data quality dimensions derived from studies found in both academic and industrial fields. It is obvious that the quantity of dimensions found in academia outnumber those in industry. This is due to the fact that academic studies of data quality dimensions were based on empirical, literature or the dimensions that could be measured objectively, not subjectively, while in industry the studies focus on specific organisational problems, and the coverage of all aspects of data quality is not their main intent (Lee et al., 2002). Accuracy is repeated in almost all studies in both fields, as correctness is commonly used to describe the accuracy concept (Weiskopf et al., 2013). Consistency concept comes second as an important dimension found in both fields collectively.

Table 2-4: Intrinsic category of dimensions from academics' and practitioners' views (Lee et al., 2002)

Academics works	Wang & Strong 1996	Zmud 1978	Jarke & Vassiliou 1997	DeLone & McLean 1992	Goodhue 1995	Ballou & Pazer 1985	Wand & Wang 1996
Accuracy	x	x	x	x	x	x	
Believability	x		x				
Completeness			x				
Consistency			x			x	
Correctness							x
Credibility			x				
Factual		x					
Freedom from bias				x			
Objectivity	x						
Precision				x			
Reliability				x	x		
Reputation	x						
Unambiguous							x

Practitioners works	**DoD** (Cykana et al. 1996)	**IRI** (Kovac et al. 1997)	**Unitech** (Mandke & Nayar n.d.)	**Diamond Technology** (Matsumura & Shouraboura 1996)	**HSBC Asset Management** (Gardyn 1997)	**AT&T and Redman** (Redman 1992)	**Vality** (Brown 1997)
Accuracy	x	x	x	x		x	
Completeness	x						
Consistency	x		x			x	
Correctness					x		
Reliability	x						

With regard to dimensions that fall into the contextual category shown in Table 2.4, the completeness dimension is vital in both fields since it is used in many studies as an aspect of quality. It is worth mentioning that completeness is regarded as contextual or intrinsic dimension, depending on the definition of completeness used in the study. Timeliness and relevance are considered crucial dimensions.

Table 2-5: Contextual category of dimensions from academics' and practitioners' views (Lee et al., 2002)

Academics works	Wang & Strong 1996	Zmud 1978	Jarke & Vassiliou 1997	DeLone & McLean 1992	Goodhue 1995	Ballou & Pazer 1985	Wand & Wang 1996
Appropriate amount	x						
Completeness	x			x		x	x
Content				x			
Currency				x	x		
Importance				x			
Informativeness				x			
Level of details					x		
Non-volatility			x				
Quantity		x					
Relevance	x		x	x			
Reliable/timely		x					
Secure currency			x				
Sufficiency				x			
Timeliness	x		x	x		x	
Usage			x				
Usefulness				x			
Value-added	x						
Practitioners works	**DoD** (Cykana et al. 1996)	**IRI** (Kovac et al. 1997)	**Unitech** (Mandke & Nayar n.d.)	**Diamond Technology** (Matsumura & Shouraboura 1996)	**HSBC Asset Management** (Gardyn 1997)	**AT&T and Redman** (Redman 1992)	**Vality** (Brown 1997)
Attribute granularity						x	
Completeness			x		x	x	
Comprehensivenes s						x	
Currency					x	x	
Essentialness						x	
Relevance						x	
Timeliness	x	x	x				

All dimensions of representational category occurred once, apart from conciseness, interpretability, meaningfulness, readability and understandability. Consistency and interpretability were found in both fields, indicating that they attract academics' and practitioners' attention.

Table 2-6: Representational category of dimensions from academics' and practitioners' views (Lee et al., 2002)

Academics works	Wang & Strong 1996	Zmud 1978	Jarke & Vassiliou 1997	DeLone & McLean 1992	Goodhue 1995	Ballou & Pazer 1985	Wand & Wang 1996
Aliases			x				
Appearance				x			
Arrangement		x					
Clarity				x			
Comparability				x			
Compatibility					x		
Conciseness	x			x			
Consistent	x						
Format				x			
Interpretability	x		x				
Lack of Confusion					x		
Meaningfulness					x		x
Origin			x				
Presentation					x		
Readability		x		x			
Reasonable		x					
Semantics			x				
Syntax			x				
Understandability	x			x			
Uniqueness				x			
Version control			x				
Practitioners works	**DoD** (Cykana et al. 1996)	**IRI** (Kovac et al. 1997)	**Unitech** (Mandke & Nayar n.d.)	**Diamond Technology** (Matsumura & Shouraboura 1996)	**HSBC Asset Management** (Gardyn 1997)	**AT&T and Redman** (Redman 1992)	**Vality** (Brown 1997)
Ability to represent null values						x	
Appropriate representation						x	
Clarity of definition						x	
Consistency					x		
Efficient use of storage						x	
Format flexibility						x	
Format precision						x	
Homogeneity						x	
Identifiability						x	
Interpretability						x	
Metadata characteristics							x
Minimum unnecessary redundancy						x	
Naturalness						x	
Portability						x	
Precision of domains						x	
Representation consistency						x	
Semantic consistency						x	
Structural consistency						x	
Uniqueness	x						

As shown in Table 2.6, the accessibility dimension is recognised as important from the point of view of academics and practitioners. Security is considered an important aspect of data quality in both fields. In academics' view, ease of use is an important data quality dimension.

Table 2-7: Accessibility category of dimensions from academics' and practitioners' views (Lee et al., 2002)

Academics works	Wang & Strong 1996	Zmud 1978	Jarke & Vassiliou 1997	DeLone & McLean 1992	Goodhue 1995	Ballou & Pazer 1985	Wand & Wang 1996
Accessibility	x		x	x	x		
Assistance					x		
Ease of Use	x				x		
Locatability					x		
Privileges			x				
Quantitativeness				x			
Security	x						
System availability			x				
Transaction availability			x				
Usableness				x			
Practitioners works	**DoD** (Cykana et al. 1996)	**IRI** (Kovac et al. 1997)	**Unitech** (Mandke & Nayar n.d.)	**Diamond Technology** (Matsumura & Shouraboura 1996)	**HSBC Asset Management** (Gardyn 1997)	**AT&T and Redman** (Redman 1992)	**Vality** (Brown 1997)
Accessibility				x	x		
Flexibility						x	
Obtainability						x	
Privacy			x				
Reliability		x					
Robustness						x	
Security			x				

To sum up, there is general agreement on neither data quality dimensions nor their definitions (Richard Y. Wang et al., 1995). Literature provides a number of classifications of data quality dimensions. However, the definitions of most differ from one classification to another. These discrepancies are because of the contextual nature of the quality (Batini et al., 2009). For example, the 'accuracy dimension' is considered one of the most frequent dimensions in data quality methodologies and classifications. It is defined by Wang & Strong (1996) as 'the extent to which data are correct, reliable and certified', while Ballou & Pazer (1985) consider data to be accurate when data values stored correspond to real-world values. The discrepancies are seen clearly when we study the definitions of 'completeness dimension' of the classifications provided in the literature. Table 2.7 shows that there are many different definitions provided for completeness in the literature. This is mainly because of the context to which each study refers, for example, information systems in Naumann (2002), Wand and Wang (1996) and data warehouse in Jarke et al. (2013).

Table 2-8: Discrepancies in the definitions of completeness provided in the literature

Reference	definition
Wang & Strong 1996	Extent to which data are of sufficient breadth, depth, and scope for the task at hand
Jarke et al. 2013	The percentage of the information stored in the sources with respect to the information in the real world
Naumann 2002	A combination of two criteria; coverage and density. Ratio between the number of non-null values in a source and the size of the universal relation
Wand & Wang 1996	Ability of an information system to represent every meaningful state of a real world system

The meanings and implications of dimensions need to be well perceived and understood by everyone in an organisation (Wand & Wang, 1996). Therefore, a consistent definition of each dimension would ease the process of developing a framework for data quality assessment. Also, data consumers may prioritise these dimensions according to their importance to the task at hand.

2.2.5 Total data quality management

The total data quality management (TDQM) programme hosted and officially launched by the MIT university laid a foundation for data quality research (Zhu et al., 2014) and developed the TDQM framework. TDQM framework extends the Total Quality Management (TQM) to the domain of data, and adapts the widely used Deming Quality Cycle (Edwards Deming, 1982). The framework advocates continuous data quality improvement by following a continual cycle of the four components of the framework:

- Define: the data consumers define the data quality in terms of *fitness for use* and identify relevant dimensions of data quality;
- Measure: the measures need to introduced for the measurement of data collections;
- Analyse: in this step, the measurement results are interpreted, level of data quality is specified and priorities are set;
- Improve: this step implements improvement initiatives.

The definition stage of the TDQM cycle identifies relevant dimensions as well as the corresponding data quality requirements. Subsequently, the measures associated with each dimension need to be developed and produced to recognise the status of data quality. With the analysis component, the root causes of data quality and the impact of poor quality data are identified. The improvement component provides techniques and steps for improving data quality.

2.2.6 Poor data quality and its impact

It is evident that data of poor quality have become critical issues facing organisations (Wand & Wang, 1996; Redman, 1998). The importance of high data quality in any organisation needs to be perceived as essential to the success of the organisation. Data with a low level of quality would normally lead to negative effects on business, not only through financial matters but the loss of trust from valuable customers. Redman (1996) states that poor data quality affects operational, strategic and tactical decisions. Poor data quality also is seen as a cause for missing potential new business, an increase in operational costs and a loss of motivation in employees (Leitheiser, 2001).

Poor data quality can adversely affect patient health and safety. A review of the safety and quality of healthcare in the United States recognises medical errors as causes of between 44,000 and 98,000 deaths in the US each year (Institute of Medicine, 2000). Not all deaths were linked primarily to poor data quality, however the report emphasises that accurate and timely patient information available at the point of care will improve patient safety.

The impact of poor data quality on an enterprise could be any of the following different levels categorised by Redman (1998):

- Impact on operations: poor quality of data results in increased cost, customer dissatisfaction and decreased employee job satisfaction. The cost resulting from the customer service organisation to correct customer details used to send to or inform the customer about current deals is a typical example of increased cost. Customers expect their names and addresses to be correct, and having deliveries sent to wrong addresses is simply not forgivable. Finally, a very high level of morale would not be expected from a hotel receptionist dealing with an exhausted traveller whose reservations have been lost.
- Impacts at the tactical level: the decision-making process is affected and compromised by poor data quality. The slightest suspicion of poor data often hinders managerial seniors from taking decisions.
- Strategic impacts: as a strategy is a decision-making process, strategy making is adversely affected by poor data quality. The impact on strategic level is great as strategy has long-term consequences to organisations. When an organisation's strategy is rolled out, plans are

deployed and results are to arrive. If these results are in poor quality, difficulty in the execution of the strategy is to be expected.

2.2.7 Data quality frameworks

A data quality framework is perceived and defined as a tool that helps an organisation to assess data quality within the organisation (Wang & Strong, 1996). Meyen & Willshire (1997) went beyond that and developed a framework not just to define a model of data, identify relevant data quality attributes and analyse data quality in their context, but to provide a guide for data quality improvement. Moreover, Eppler & Wittig (2000) stated some goals that a framework needs to achieve. They emphasised that data quality problems should be analysed and solved through scheme provided by the framework. The framework should provide the basis for proactive management. Wang and Strong (1996), Meyen and Willshire (1997) and Eppler and Wittig (2000) have developed frameworks to assess and review systems with regard to data quality issues within organisations.

The most widely used category for data quality dimensions is the one introduced in the data quality frameworks developed by Wang & Strong (1996). Based on limited literature and their experience with data consumers, they proposed a preliminary conceptual framework classified into accuracy, relevancy, representation and accessibility. However, this labelling did not capture the essence of the underlying dimensions as a group. That is, the group of dimensions that was labelled accuracy, for example, was richer than what is conveyed by the label of 'accuracy'. Therefore the categories' labelling was later altered to intrinsic DQ, contextual DQ, representational DQ and accessibility.

2.2.8 Data Quality Assessment

Data quality assessment is a hard and unending task, even if there is certainty about what is to be measured, and it is considered a complex task. The literature has been enriched by many studies in order to improve data quality in organisations (Wang & Strong, 1996; Strong et al., 1997; Ballou & Pazer, 1985; Batini et al., 2009; Wang, 1998). In order for an organisation to assess its organisational data quality and improve it to facilitate decision-making process, it needs to be able to measure and assess its data. Any data quality activity needs to have an overall data model as well as an associated assessment instrument for measuring data quality. It is defined as 'the process of assigning numerical or categorical values to quality criteria in a given setting' (Gertz et al., 2004).

It is accepted that data quality cannot be assessed independently of data consumers (Strong et al., 1997). The same data used by different tasks or different users may require different level of quality. For example, misspellings can be tolerable in one occasion but not another. To have data quality assessed, subjective and objective data quality measures associated with each dimension must be considered (Ge & Helfert, 2008; Pipino et al., 2002).

Subjective assessments evaluate data quality, taking the production role's views into consideration. Data quality dimensions should be developed and defined from the perspectives of data consumers (Wang & Strong, 1996). This type of assessment focuses on whether the data are fit for use from the perspective of data consumers. It takes many forms, such as questionnaires or interviews, during the assessment process to assess the defined dimensions.

With regard to objective assessments, these can be task-dependent or task-independent (Pipino et al., 2002). Task-independent measures can evaluate data without the need for the contextual knowledge of the application, and can be applied to any dataset regardless of its context. While task-dependent measures are developed for specific contexts, and usually include constraints set by the database administrators, the business rules of the organisation and regulations are provided by the government or company.

Companies should understand their needs and follow principles in order to develop the right measures for their needs. Formulating the measures is the easier part of the measures development process. The functional forms, which help formulate the measures, are simple ratio, min or max operation and weighed average (Pipino et al., 2002). *Simple ratio* measures the ratio of desired results to the total results, calculated by dividing the number of undesirable outcomes by the total outcomes and subtracting the result from one. 'One' would represent to the most desirable and 'zero' the least desirable score. This type of function is usually used with accuracy, completeness and consistency. *Min or max operation* can be applied to handle dimensions that involve the aggregation of multiple data quality variables. This type of the functional form is ideal for dimensions such as believability and the amount of data. *Weighted average* is a good functional form to involve the importance of the variable to the overall evaluation of a dimension.

Ge and Helfert (2008) state that raw data and information production are the types of data used in the assessment processes. Objective assessments mainly deal with raw data and component data, while subjective assessments deal with the final information production. They classify the data quality dimensions into two groups according to their suitable type of assessment, as shown in Figure 2.2.

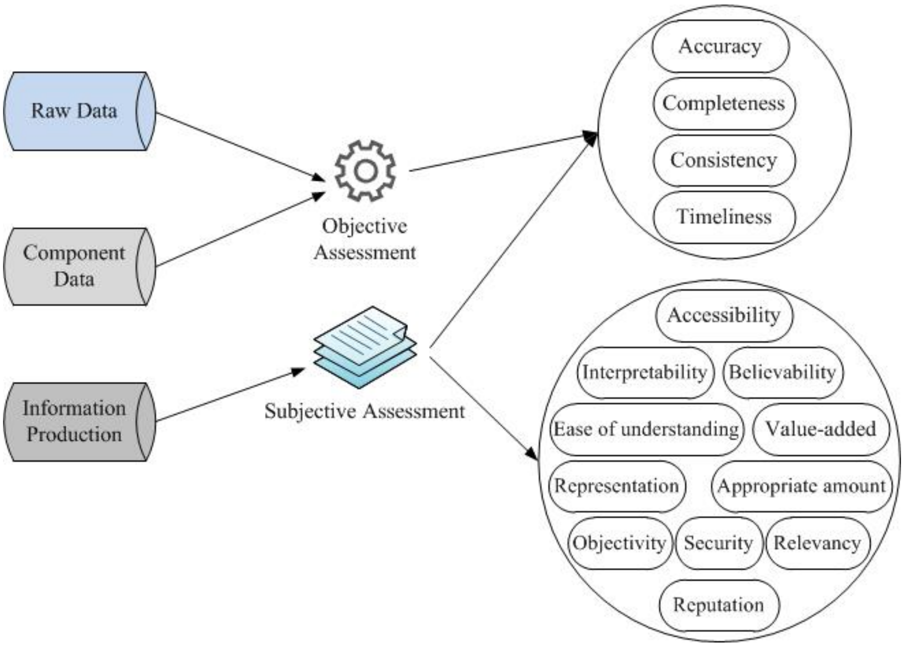


Figure 2.2: A data quality assessment model (Ge & Helfert, 2008)

Furthermore, there is a clear difference between the nature of subjective and objective assessment. Table 2.8 lists these from several aspects, including data storage, assessing results, process, criteria, and measuring object and tool.

Table 2-9: Comparison between objective and subjective data quality assessment (Ge & Helfert, 2008)

Feature	Objective assessment	Subjective assessment
Tool	Software	Survey
Measuring object	Data	Information product
Criteria	Rules, patterns	Fitness for use
Process	Automated	User involved
Assessing result	Single	Multiple
Data storage	Database	Business context

Arts et al. (2001) developed a model for data quality measurement and a framework of procedures to assure data quality. In their model, seven steps must be taken to measure data quality: (1) describe the objectives of the registry; (2) determine the data items that need to be checked; (3) define the data quality aspects; (4) select the methods for quality measurement; (5) determine the criteria; (6) perform quality measurement; (7) quantify measured data quality.

Naumann et al. (1999) propose a new method of assessing information quality by assigning numerical values to information-quality criteria used to define information quality. The criteria used have already been identified by Strong et al. (1997). Naumann and Rolker in another study (Naumann & Rolker, 2000) emphasise the difficulty of the assessment phase due to the subjective nature of assessment, quality metadata not being available, the numerous amount of data to be assessed, and information from autonomous data sources being subject to surprising changes. They provide assessment-oriented classifications of dimensions.

Some approaches to assessing information quality are based on subjective user inputs (Bobrowski et al., 1999; Lee et al., 2002; Naumann et al., 1999). Lee et al. (2002) developed a methodology for assessing information quality. It comprises three components: the PSP/IQ model, the IQ instrument and gap analysis techniques. It measures information quality on four characteristics: soundness, dependability, usefulness and usability. Each characteristic includes several information-quality criteria. As mentioned earlier, this methodology is based on subjective measures and uses a questionnaire designed to capture users' opinions for each criterion on a scale of 0–10.

In order to assess the quality of data, both subjective and objective data quality measures need to be taken into consideration (Pipino et al., 2002; Ge & Helfert, 2008). This is essential, as data quality practitioners stress that subjective and objective measurement items need to be considered in the data quality framework as they reflect the contextual nature of the data quality and their various users (Meyen & Willshire, 1997; Carson, 2000).

2.2.9 Data quality problems and dirty data

Organisations and enterprises tend to pay not enough attention to the existence of 'dirty data' in their repositories, although it compromises the quality of their data and produces unreliable information. The reasons could be due to resources, time and a lack of appreciation. In the literature, many proposals of 'dirty data' taxonomies were proposed that tackle a wide variety of data quality problems.

Muller and Freytag

In Müller and Freytag (2005), data anomalies were roughly classified into syntactical, semantic and coverage anomalies. Syntactical anomalies concern representation-related dirty data. This type includes lexical error, domain format errors and irregularities. Semantic anomalies affect the comprehensiveness of data collection as well as non-redundant representation, whilst coverage anomalies cause missing values and missing tuples. They include integrity constraint violations, contradictions and duplicates invalid tuples.

Table 2-10 shows the types of data anomalies mentioned above, with a brief description.

Table 2-10: Data anomalies identified by Muller and Freytag

Anomaly type	Definition
Lexical errors	The discrepancies between data items' structure and the specified format
Domain format errors	No conformation between the given value for an attribute and the anticipated domain format
Irregularities	Non-uniform use of values, units and abbreviations
Integrity constraint violations	It describes data that do not satisfy the integrity constraint that restrict the set of valid values
Duplicates	It shows two or more tuples representing the same entity of the mini-world
Invalid tuples	Not representing the mini-world entity due to our incomplete knowledge
Missing values	Null values for attributes where there exists NOT NULL constraint for them
Missing tuples	Tuples not being represented in the data collection

Rahm and Do

Rahm and Do (2000) provide a two-level classification of data quality problems associated with databases. In the first hierarchical model, problems are categorised as single-source and multi-source. In each, the data quality problems are classified as schema-level and instance-level problems.

With regard to the single-source category, schema-specific problems occur due to the limitations of model and application-specific integrity constraints, as the data quality of a source mainly depends on its data being governed by schema and integrity constraints. On the other hand, in multi-source category the heterogeneity of data models from different sources results in many 'dirty data' such as duplicates and instances of naming conflicts.

Table 2-11 shows all data quality problems identified by Rahm and Do (2000).

Table 2-11: Data quality problems identified by Rahm & Do

NO	Data Quality Problems	Type of problem
R1	Illegal values due to invalid domain range	**Single-source problem**
R2	Violated attribute dependences at schema level	
R3	Uniqueness violation	
R4	Referential integrity violation	
R5	Missing values	
R6	Cryptic values, Abbreviations	
R7	Misspellings	
R8	Embedded values	
R9	Misfielded values	
R10	Violated attribute dependences at instance level	
R11	Word transpositions	
R12	Duplicated records in single data source	
R13	Contradicting records in single data source	
R14	Wrong references	
R15	Naming conflicts	**Multi-source problem**
R16	Structural conflicts	
R17	Data conflicts in multiple data sources	
R18	Duplicate records in multiple data sources	
R19	Contradicting records in multiple data sources	

Kim et al.

Kim et al. (2003) take another pattern of classification of 'dirty data', looking on it as either missing data, wrong data or non-standard representations of the same data. This leads them into a hierarchically structured taxonomy; missing data, not missing but wrong, and not missing neither wrong but unusable data. Table 2-12 lists data quality problems identified by Kim et al.

Table 2-12: Data quality problems identified by Kim et al.

NO	Data quality problems	Type of problem
K1	Missing data (null value allowed)	Missing data
K2	Missing data (null value not allowed)	
K3	Use of wrong data type including value range	Not missing but wrong data
K4	Dangling data	
K5	Violation of uniqueness constraint data	
K6	Mutually inconsistent data	
K7	Lost update (lack of concurrency)	
K8	Dirty read (lack of concurrency)	
K9	Unrepeatable read (lack of concurrency)	
K10	Lost transaction (lack of concurrency)	
K11	Wrong categorical data	
K12	Outdated temporal data	
K13	Inconsistent spatial data	
K14	Erroneous entry	
K15	Misspelling	
K16	Extraneous data	
K17	Entry into wrong fields	
K18	Wrong derived-field data from stored data	
K19	Inconsistency across multiple tables/files	
K20	Different data for the same entity across multiple database	Not missing, not wrong but unusable data
K21	Ambiguous data due to use of abbreviation	
K22	Ambiguous data due to incomplete context	
K23	Different representation for non-compound data due to use of abbreviation	
K24	Different representation for non-compound data due to use of Alias/ nickname	
K25	Different representation for non-compound data due to encoding format	
K26	Different representation for non-compound data due to different representations	
K27	Different representation for non-compound data due to measurement units	
R28	Different representation for non-compound data due to abbreviation	
K29	Different representation for compound data due to use of special characters	
K30	Different representation for compound data due to different ordering	
K31	Different representation for hierarchical data due to abbreviation	
K32	Different representation for hierarchical data due to use of special characters	
K33	Different representation for hierarchical data due to different ordering	

Oliveira et al.

This is a recent work compared to those above. Researchers Oliveira and Rodrigues (2005) present a comprehensive taxonomy of data quality problems by reviewing previous work (Kim et al., 2003; Rahm & Do, 2000; M□ller & Freytag, 2005), using a bottom-up approach from the lowest to the highest levels at which data quality problems appear. This results in six levels of granularity, ranging from problems in single attribute value (lowest level) to problems in multi-source (highest level), as shown Table 2.12.

Table 2-13: Data quality problems from Oliveira and Rodrigues

NO	Data Quality Problems	Level
O1	Missing value	**An attribute value of a single tuple**
O2	Syntax violation	
O3	Outdated value	
O4	Interval violation	
O5	Set violation	
O6	Misspelled error	
O7	Inadequate value to the attribute context	
O8	Value items beyond the attribute context	
O9	Meaningless value	
O10	Value with imprecise or doubtful meaning	
O11	Domain constraint violation	
O12	Uniqueness value violation	**Values of a single attribute**
O13	Synonyms existence	
O14	Domain constraint violation	
O15	Semi-empty tuple	**The attribute values of a single tuple**
O16	Inconsistency among attribute values	
O17	Domain constraint violation	
O18	Redundancy about an entity in single data source	**Attribute values of several tuples**
O19	Inconsistency about an entity in single data source	
O20	Domain constraint violation	
O21	Referential integrity violation	**Relationships among multiple Relations**
O22	Outdated reference	
O23	Syntax inconsistency in single data source	
O24	Inconsistency among related attribute values	
O25	Circularity among tuples in a self-relationship	
O26	Domain constraint violation	
O27	Syntax inconsistency in multi data sources	**Multiple data sources**
O28	Different measure units in multi data sources	
O29	Representation inconsistency in multi data sources	
O30	Different aggregation levels in multi data sources	
O31	Synonyms existence in multi data sources	
O32	Homonyms existence	
O33	Redundancy about an entity in multi data sources	
O34	Inconsistency about an entity in multi data sources	
O35	Domain constraint violation in multi data sources	

Müller and Freytag's (2005) classification of data anomalies does not cover as many as data quality problems as the rest. This is because they do not take problems from multi- data sources into consideration. Rahm and Do's (2000) classification of data quality problems is widely used and much cited. It considers the problems in both single and multi-data sources. Kim et al. (2003) and Oliveira et al. (2005) proposed more dirty data types than others.

2.2.9.1 Data Quality AND Dirty Data Taxonomies

In the literature, data quality problems and 'dirty data' are used interchangeably, addressing issues and problems that lead to poor quality of data and, consequently, produce unreliable data. Müller and Freytag (2005) used the term 'data anomaly' instead, and classified these anomalies as lexical errors, domain format errors, irregularities, integrity constraint violations, duplicates, invalid tuples, missing values and missing tuples.

Others (Rahm & Do, 2000; Kim et al., 2003; Oliveira et al., 2005) consider data quality problems as occurring multi-source, in contrast to Muller and Freytag (2005). Rahm and Do group their findings into multi-source and single-source problems. Their classification has been the widest and most cited in the context of data cleansing. Kim et al. (2003) produced from their research a comprehensive hierarchal taxonomy that captures 33 types of 'dirty data', both single and multi-source. In their approach, they rely on the fact that 'dirty data' is either missing, wrong or unusable. The most recent proposal is from Oliveira and Rodrigues (2005), who adopt a bottom-up approach to generate 35 types of 'dirty data'.

2.3 Data Quality in Health Context

The healthcare field is known to be information-intensive, since massive data information is generated on a daily basis. An estimated 30 per cent of the health budget usually goes on issues related to information handling (Health Information and Quality Authority, 2011). Data quality issues are of critical concern in the healthcare sector for several reasons: life or death decisions rely on the accurate information; and the quality of health data is known to be highly variable and often incorrect (Raghupathi & Raghupathi, 2014). Sound, accurate and reliable health information plays an important role in providing safe and reliable healthcare. Similarly, this high quality of information helps decision makers in their healthcare planning.

2.3.1 Data quality improvement in healthcare

The report of the Institute of Medicine (2000) on medical errors and the poor quality of care awakened the health care system to the challenge of decreasing the number of medical errors and adverting incidents in hospitals. Many healthcare providers and organisations implemented total quality management (TQM) practices in order to improve the care delivery and patient safety. For instance, National Health Service (NHS) in the United Kingdom has launched the 'NHS Modernisation Programme' through which many changes were implemented to healthcare management and provision at local, regional and national level. 'Clinical pathways' are a clear example of quality management tools that reduce the variability in clinical practice and improve outcomes. They detect if patient data vary from the expected outcomes. This means that the quality of data is paramount to organisations. Therefore, high-quality data is essentially important to any TQM programme. Total data quality management

(TDQM) adapted the philosophy and discipline of TQM to allow data quality improvements (Wang & Strong, 1996).

The implications of poor quality of data in the healthcare sector point out the essence of data quality management, and motivate professionals and managers to plan and implement strategies for data quality. The following examples show how single errors can lead to significant effects on patients (Rigby et al., 2001):

- Errors in updated embedded clinical coding software giving false plain language representation of diagnoses, United Kingdom.
- Errors in reference database calculation of Down's syndrome screening, giving false negatives.
- Age cohort of women omitted from call up for cervical screening, Grampian Region, Scotland.
- Error in software calculating risk of Down's syndrome led to falsely low calculation of risk for 150 women, Sheffield.

The importance of high-quality health information for healthcare decisions is widely recognised by many health-related bodies through their attempts and initiatives (Batini et al., 2009). For instance, Article 6 of the European Union data protection directive addresses data quality and emphasises that data must be accurate, up-to-date, processes fairly and lawful, relevant and not excessive (Council of the european union, 1995). Lack of trust in the quality of health data is a major barrier to the effectiveness of use of data populating health information systems (van der Veer et al., 2010). Data quality control is essential to support and help quality improvement projects in the healthcare sector (Needham et al., 2009).

Improving data quality in healthcare settings is a difficult and complex task. This is due to the complexity of healthcare, as well as different perspectives on data quality in the same dataset (Cabitza & Batini, 2016). Multiple uses of the same data from multiple users make the data quality improvement even harder. The nature of data collection in healthcare imposes the fact that medical records and registries tend to be incomplete (Shortliffe & Barnett, 2001). A review conducted by Lorence and Jameson (2002) on data quality practices in US healthcare organisations showed that there was regional variation in such practices that hinder the idea of comparisons and benchmarking with broader health sector data. Common standardised practices for data quality assessment are needed in order to make national comparisons of healthcare data.

2.3.2 Data quality improvement in the EHR

Given the fact that EHRs surpass the performance of many existing data repositories, and that there is growing interest in the development and implementation of EHRs by healthcare bodies (Weiskopf & Weng, 2013; Botsis et al., 2010), it is unsurprising to learn that many studies have been undertaken to assess and improve data quality in EHR systems. Orfanidis et al. (2004) found that the analysis and design of data quality issues are an integral part of the development and implementation of EHR systems and should be addressed from the inception of the project.

Weiskopf and Weng (2013) reviewed the methods and dimensions of data quality assessment for EHR data. They found 27 unique terms describing the dimensions of data quality in the related literature. Based on the definitions of these terms, the lists of the terms were mapped to five dimensions. The five empirically derived data quality dimensions are completeness, correctness, concordance, plausibility and currency. They felt that the correctness (accuracy), completeness and currency (timeliness) are fundamental to and core concepts of data quality in EHR context, and considered them to be non-reducible. The remaining two – concordance (consistency) and plausibility – were seen in their study as proxies to serve the fundamental dimensions.

The Canadian Institute for Health Information (CIHI) defined data quality in the context of users; that is, if data satisfies users' needs, then they are fit for use (Canadian Institute for Health Information, 2009). The dimensions of data quality identified are accuracy, timeliness, comparability, usability and relevance.

2.3.3 Data Quality Dimensions

As discussed earlier, the quality of data may be determined through assessment against a set of dimensions. The clinical research community has failed to develop a consistent taxonomy of data quality as there is an overlap of terms among existing dimensions (Weiskopf & Weng, 2013). As can be seen in Table 2-14, many researchers have defined data quality dimensions in the context of health. Weiskopf and Weng (2013) concluded in their review study that the dimensions used in data quality assessment process of EHR data are completeness, correctness, concordance, plausibility and currency, while the Canadian Institute for Health Information (CIHI) developed a data quality framework in which the health data are assessed in a common and objective approach. The framework is structured along five dimensions of data quality: accuracy, timeliness, comparability, usability and relevance. These dimensions are divided into 24 characteristics, and are further made up 86 criteria.

Evaluation is performed through questionnaires scoring each criterion on a four-point scale as ‘not applicable’, ‘unknown’, ‘not met’ and ‘met’.

Table 2-14. Data quality dimensions in health information systems

Research	Data Quality Dimensions
Canadian Institute for Health Information, 2009	Accuracy, timeliness, comparability, usability, relevance
Liaw et al., 2012	Completeness, consistency, correctness, timeliness, Relevance, usability, security
Weiskopf & Weng, 2013	Completeness, correctness, concordance, plausibility, currency.
Kerr et al., 2007	Accuracy, timeliness, comparability, usability, relevance, privacy, security

Kerr et al. (2007) developed a data quality framework for the New Zealand health sector to provide a consistent assessment tool across the national health information databases and EHRs. The framework was modelled on the findings of the CIHI. With the inclusion of data quality dimensions introduced by CIHI, the framework introduced privacy and security.

As can be seen from the literature, the most common dimensions of data quality are accuracy, completeness, consistency and timeliness (Batini et al., 2009; Liaw et al., 2012). These dimensions should be the founding set for data quality methodologies as they reflect the most important facets.

2.3.4 International reviews of data quality strategies in healthcare

Government departments of health have recognised the need for data quality practice in order to have accurate, complete, reliable, relevant and available data in a timely manner that help and support decision-makers for healthcare delivery. Many national initiatives and strategies, such as Data Quality Maturity Index (DQMI) in UK, have been proposed to prevent data quality problems, and to manage data proactively.

United Kingdom

There are many numbers of programmes within UK that indicate that data quality is seen as essential to enable the delivery of safe care. These programmes are to improve and monitor data quality, to enhance the information infrastructure, and to develop information standards. Many resources were dedicated to support the initiatives in UK, namely:

- NHS Connecting for Health:[3] a Directorate of the Department of Health, established in 2005, whose main goal is to deliver the National Programme for Information Technology.
- Information Governance Toolkit[4]: a web-based application developed to allow organisations to self-assess the way that they process or handle information.
- NHS digital[5]: the Information Centre for Health and Social Care is an authority responsible for providing information, data and IT systems for commissioners, analysts and clinicians in health and social care. It has established a Data Quality service to support improvements in data quality. The Data Quality Maturity Index (DQMI) is a quarterly publication to promote the significance of data quality in the NHS by providing data owners with information about their data quality.
- Secondary Uses Service: provides services to support the analysis of data collected during the provision of treatment for patients. The quality of data collected from secondary care providers are assessed to assure the high quality of data. (Health Information and Quality Authority, 2011)

Canada

In Canada, there have been many data quality initiatives to provide support to healthcare organisations to improve data quality. The Canadian Institute for Health Information (CIHI)[6] is a non-profit, independent organisation dedicated to providing essential information on Canadian's health systems and the health of Canadians. It is responsible for a number of databases and registries that capture information during the care process. CIHI provide publications on information on health indicators to monitor and compare performance. With regard to data quality improvement, CIHI has several initiatives, including the following:

- Data Quality Framework: it was developed for improving data quality in CIHI. The framework consists of three main components; a data quality work cycle, a data quality assessment tool and documentation about data quality.
- Data Quality in Action: it is meant to conduct a range of data quality activities. These activities provide education to data providers and coding query service for users of ICD-10-CA/CCI, and publish standards for financial reporting and ICD-10-CA/CCI.

[3] http://www.connectingforhealth.nhs.uk/
[4] https://www.igt.hscic.gov.uk/
[5] http://content.digital.nhs.uk/
[6] https://www.cihi.ca/en

- Infostructure Standards: CIHI publishes health infostructure standards that comprise technical specifications that ease and facilitate the interoperability of communication and information technology. Standardisation of the collection and sharing of information contributes to data quality.

New Zealand

The National Health IT Board[7] (NHITB) is sub-committee of the National Health Board (NHB), and provides strategic advice to the Ministry of Health. It is supported by a number of advisory groups such as the Health Information Standards Organisations (HISO). HISO improve the New Zealand health system by supporting and promoting the development and understanding health information standards. Moreover, data submitted to national collections are checked to assure high quality and, in the case of bad quality data, data providers must correct and resubmit data.

2.4 Instrument practicality survey

This Section aims to review and discover the factors leading to individual's adoption of an instrument. These factors would be featured on a questionnaire to understand the practicality and applicability of the proposed DQ instrument, and examine their post-adoption environment where individuals decide between continuing and discontinuing usage of the proposed instrument. Thus, factors that measure the acceptability and continuance of the instrument are highlighted. The Technology Acceptance Model (TAM) and Expectation-Confirmation Model (ECM) are investigated to come up with the determinants that examine the practicality of the data quality instrument derived from the data quality framework.

2.4.1 Technology Acceptance Model (TAM)

TAM model was developed by Davis (1985), based on the Theory of Reasoned Action (TRA). It is one of the most referenced models in the adoption of information technology (Venkatesh, 1999). The purpose of this model is to predict the acceptability of an information system.

[7] http://healthitboard.health.govt.nz/

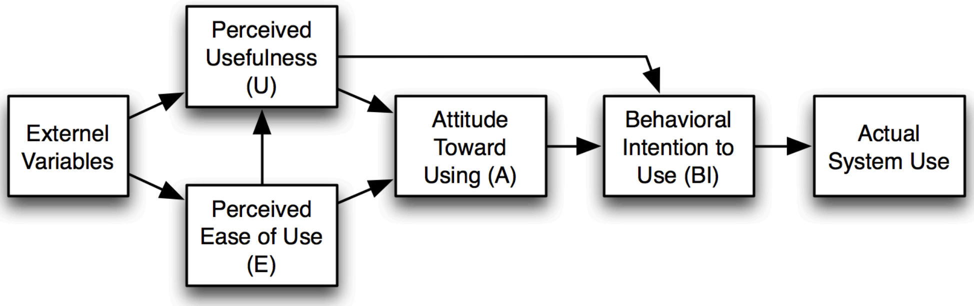

Externel Variables
Perceived Usefulness (U)
Perceived Ease of Use (E)
Attitude Toward Using (A)
Behavioral Intention to Use (BI)
Actual System Use

Figure 2.3. Technology Acceptance Model (Davis et al., 1989)

This model suggests that there are two salient determinants that impact the user's attitude towards systems acceptability: perceived usefulness and perceived ease of use. Perceived usefulness refers to an individual's subjectively assessed probability that using a particular system will enhance their job performance. Perceived ease of use refers to the degree to which an individual expects that the use of that system does not require more effort (Davis et al., 1989).

A number of studies have applied the TAM model over the years to predict the acceptability of IT in healthcare (Hu et al., 1999; Yi et al., 2006). More recently, many studies have employed additional constructs in the model to consider the context of IT in healthcare (Klein, 2007; Liu & Ma, 2005; Ilie et al., 2009; Moores, 2012).

However, one of the criticisms of the TAM model is that it lacks any explanation of the behavioural and performance-based consequences of adoption. It only highlights the beliefs that influence a user's attitude towards the system. This indicates that more factors could work in combination with the identified factors to predict the adoption and the continuance. The Expectation-Confirmation Model was developed to draw factors that influence a user's intention to the continuance of their use of an instrument.

2.4.2 Expectation-Confirmation Model

Bhattacherjee (2001) adapted the Expectation-Confirmation Theory (ECT) of consumer behaviour research by integrating it with prior IS use literature to identify the determinants of IS continuance intention. It addresses the factors that influence users' attitudes towards the continuance of use. As seen in Figure 2.4, perceived usefulness and user satisfaction with prior use influence IS continuance intention. User satisfaction is primarily determined by users' confirmation of expectation from prior use and secondarily on perceived usefulness. It also shows the significant influence of confirmation on perceived usefulness.

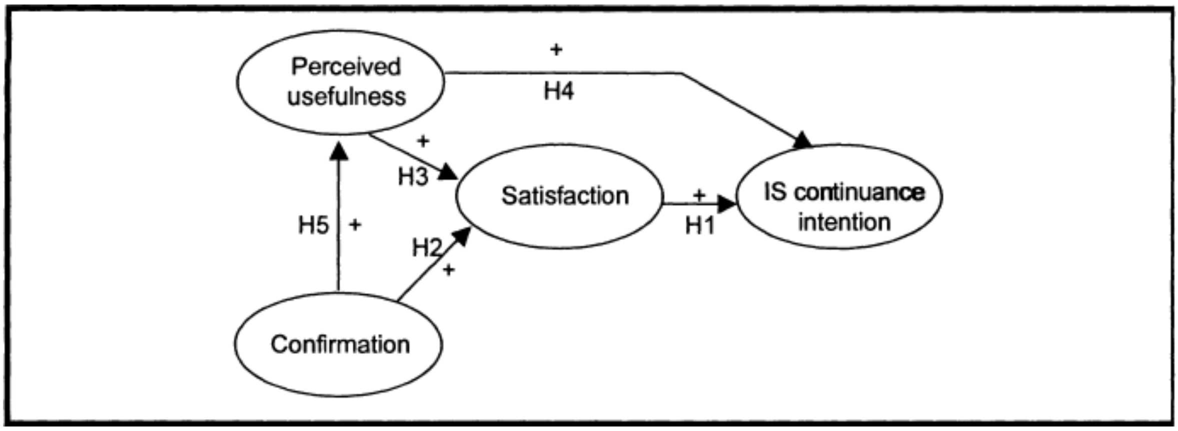


Figure 2.4: Expectation-Confirmation Model of IS continuance (Bhattacherjee, 2001)

2.4.3 Synthesis of influential factors

Thong et al. (2006) have expanded the Expectation-Confirmation Model (ECM), and included the perceived ease of use adapted from TAM. The post-adoption belief, the perceived ease of use, was empirically tested and contributes to the development of a more comprehensive account of IS continuance behaviour.

Perceived usefulness measures the instrumentality of the instrument, and perceived ease of use taps into self-efficacy. These two constructs influence the continuance decision, as they are the key motivators of the adoption (Bhattacherjee, 2001).

The other two constructs are satisfaction and confirmation. Satisfaction refers to a user's affect with prior instrument use, while confirmation refers to the users' perception of congruence between expectation of the DQ instrument use and its actual performance (Bhattacherjee, 2001).

2.5 Discussion

Few publications have addressed data quality in healthcare systems (Kahn et al., 2012). Existing frameworks of data quality dimensions have been based on literature reviews, industry experience or intuitive understanding. The definition of a dimension may vary from one framework to another – see the example given by Wand and Wang (1996) in their definition of accuracy. The concept of data quality depends on the actual use of the data. Thus, it depends on the application: what is considered high quality data in one application may not be sufficient in another (Wand & Wang, 1996). Wand and Wang (1996) also emphasise the importance of providing a design-oriented definition of data quality that will reflect the nature of information systems.

2.6 Chapter Summary

As mentioned earlier, data quality is a multidimensional concept, and there is no consensus on a rigorously defined set of data quality dimensions. This is due to the definition of quality of data being 'fitness for use'.

EHRs are seen as a promising solution to problems in health information management, despite threats posed during data storage and transmission. However, the key barrier to optimal use of routinely collected data is an increasing quantity of data exhibiting poor quality. This could boost the need to automate both data quality measurement and semantic interoperability (Liaw et al., 2012).

Besides supporting clinical care, EHRs with sufficient data quality are being used for secondary research and evaluation purposes (Taggart et al., 2012).

The assessment phase of data quality methodology is a difficult task, and needs to consider both subjective and objective data quality measures. This is due to the fact that some dimensions, such as interpretability and relevance, cannot be objectively assessed.

Chapter 3 Data Quality Framework for EHR

Data quality was discussed in previous chapters, as this challenge greatly affects the quality of care in the healthcare sector. Although the nature of data in EHRs allows data to be exchanged and integrated, few studies have addressed data quality in the healthcare sector (Kahn et al., 2012). The absence of a rigid data quality framework to assess data in EHRs could lead to poor data quality as, for example, provenance has not been much addressed as a dimension of data quality in the context of health. The research began by identifying the data quality dimensions related to EHR context to answer the first question in a data quality framework development.

> **RQ1:** What data quality determinants are important for EHR stakeholders perceived data quality?
>
> **RQ2**: Based on the proposed framework, what is the appropriate instrument with which to measure quality of data in EHRs?

From the literature review in the previous chapter, a basic set of data quality dimensions could be defined as accuracy, completeness, consistency and timeliness; these dimensions reflect those chosen by majority of researchers (Batini et al., 2009; Liaw et al., 2012). This study has developed an EHR-specific framework to cover all data quality aspects that need to be assessed in the EHR context. First, all data quality dimensions proposed in clinical research were gathered and normalised using data quality characteristics that are widely accepted in the literature review.

3.1 The Proposed Data Quality Framework

A framework of data quality was developed to answer research question RQ1. The development of this framework went through many stages, and Figure 3.1 shows the process. It began by gathering the dimensions of data quality in the EHR context, and then mapping them into the basic set of data quality dimensions of most frameworks: that is, accuracy, completeness, consistency and timeliness. This step was to benefit the widely recognised data quality frameworks that target different models of data.

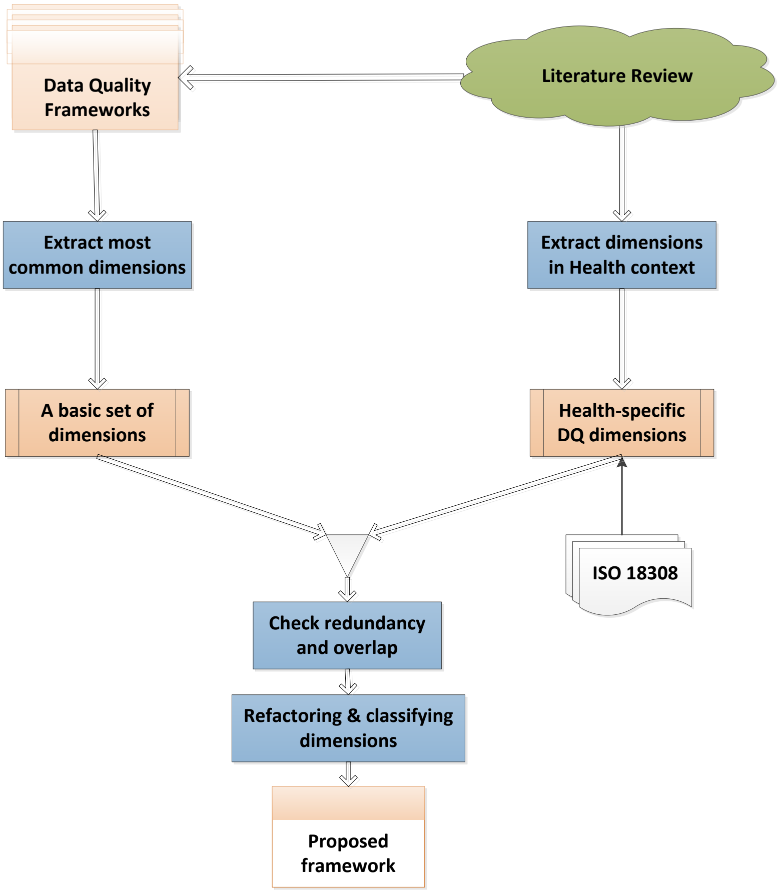

Data Quality Frameworks
Literature Review
Extract most common dimensions
Extract dimensions in Health context
A basic set of dimensions
Health-specific DQ dimensions
ISO 18308
Check redundancy and overlap
Refactoring & classifying dimensions
Proposed framework

Figure 3.1: The framework development process

3.1.1 Framework development process

As discussed in he was not happy

2.3.3, data quality dimensions introduced to the health context were pooled and normalised in order to develop a set distinct dimensions. This includes the work of Weiskopf and Weng in 2013, in which the researchers reviewed the clinical literature and identified 27 terms describing dimensions of data quality, from which they derived five dimensions (Weiskopf & Weng, 2013). It also includes the framework introduced by CIHI (Canadian Institute for Health Information, 2009), Kerr et al. (2007) and Liaw et al. (2012) (see Figure 3.2).

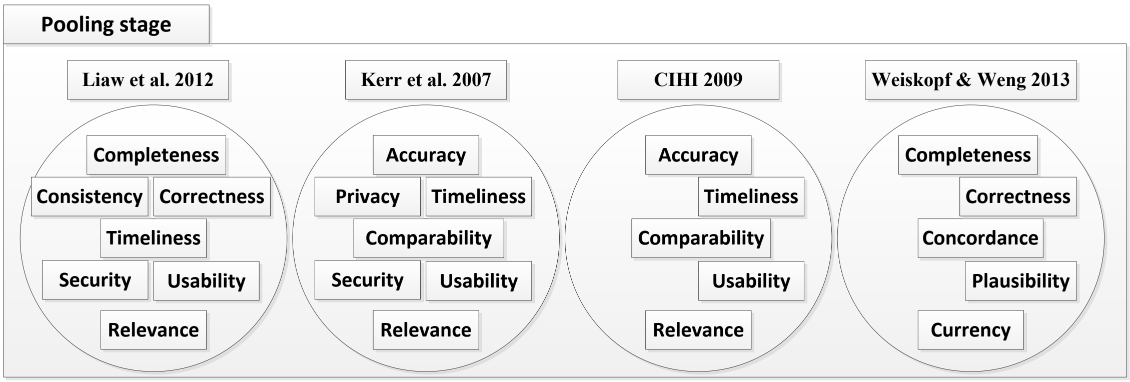


Figure 3.2: Pooling stage of dimensions of data quality

There was a clear variability and overlap in the dimensions extracted from the studies mentioned earlier. For example, accuracy was sometimes used as a synonym for correctness, but in other sources meant correctness as well as completeness (Weiskopf & Weng, 2013). In order to abstract the dimensions in such a way as to be exhaustive, the widely adopted data quality framework (Wang & Strong, 1996) is used as a reference to facilitate choosing the right term for a dimension.

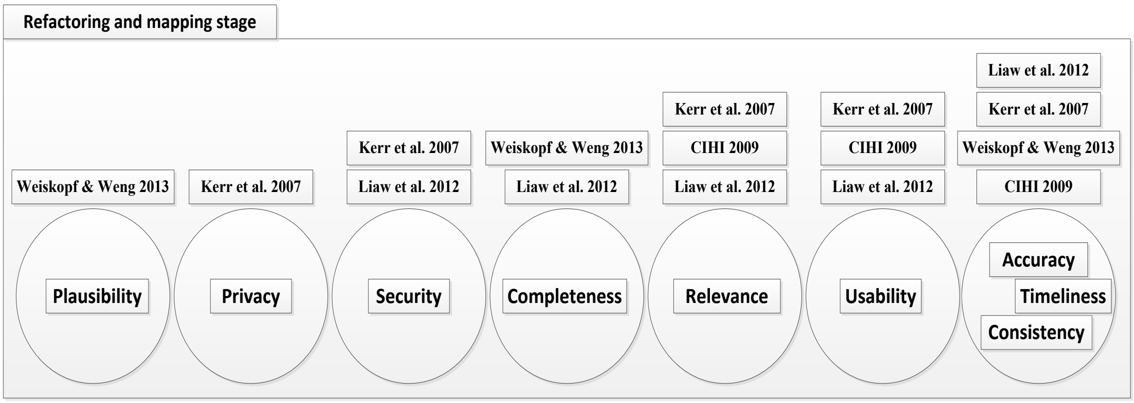


Figure 3.3: Refactoring and mapping stage of dimensions of data quality

Accuracy, timeliness and consistency

As seen in Figure 3.3, all frameworks and studies confirmed the fact that 'accuracy' ('correctness') is a very important characteristic of data quality in EHRs and the healthcare context. 'Accuracy' and 'correctness' are used interchangeably to mean that data values stored in systems correspond to real-world values. However, 'accuracy' is a broader term as it refers to 'the extent to which data are correct, reliable and certified' (Wang & Strong, 1996). Moreover, 'accuracy' is the most frequently used of similar terms in data quality literature in general (see Section 2.2.4).

With regard to 'timeliness', there are many terms that describe or similar to 'timeliness' or its implication in the literature, such as currency, age and volatility. 'Currency' or 'age' covers just part of the time-related aspect of data quality, as it implies the degree to which data is up-to-date (Redman, 1996), while volatility describes the time period for which data are valid (Jarke & Vassiliou, 1997). On the other hand, 'currency' and 'volatility' are considered two components of 'timeliness' (Bovee et al., 2003).

Consistency is an important dimension of data quality, recognised by the selected frameworks of data quality in healthcare context. Besides, consistency is among the basic set of data quality dimensions and captures the attention of the majority of researchers (Batini et al., 2009). It is worth noting that there is clear variability in terms of dimension names, since the Canadian Institute for Health Information (2009) and Kerr et al. (2007) chose 'comparability', Weiskopf and Weng (2013) named this 'concordance' and 'consistency' was picked by Liaw et al. (2012). However, the Canadian Institute for Health Information (2009) and Kerr et al. (2007) clearly defined 'comparability' as 'the extent to which data holdings are consistent over time', and Weiskopf & Weng (2013) pointed out that 'consistency' is a synonym of 'concordance'. Thus, 'consistency' is a proper name for this aspect of data quality, due to its wide use in the context of data quality frameworks.

Usability and relevance

Usability is an important aspect of data quality, and one of the cognitive challenges that deters the potential of EHRs (Zhang & Walji, 2011; Wand & Wang, 1996). It is defined as the ease with which data can be accessed, used, updated, understood, maintained and managed. If information products are difficult to use, they can be rendered worthless, regardless how accurate, timely and complete they may be (Canadian Institute for Health Information, 2009; Liaw et al., 2012).

'Relevance' of data is another characteristic that appears in many frameworks of data quality. It is the extent to which information is appropriate and useful for the intended task. It is a very important constituent of data quality that tells whether the available data or information products inform the issues most important to stakeholders. The 'relevance' dimension reflects whether the data available for the task at hand are valuable to their users.

Completeness

'Completeness' is one of the most frequently used dimension of data quality in healthcare (Weiskopf & Weng, 2013; Liaw et al., 2012). It is the state in which information is not missing and is sufficient for the task. It is also the most common term used among its similar terms. It is one of the four attributes of data quality relevant to patient records, as identified by the Institute of Medicine (Weiskopf & Weng, 2013). Moreover, the literature of data quality suggests that the core dimensions of the data quality framework encompass 'accuracy', 'completeness', 'consistency' and 'timeliness' (Liaw et al., 2012; Batini et al., 2009). Despite the fact that completeness is not considered as a dimension by the Canadian Institute for Health Information (2009) or Kerr et al. (2007), they both used it as an essential criterion within their proposed frameworks.

Security and privacy

The crucial and sensitive content of data residing in EHR systems emphasises that security and privacy are vital requirements to their management. They are the factors that contribute to data stability. Data security is aimed at protecting patient and personal data from unauthorised access or disclosure (Van Der Haak et al., 2003). On the other hand, privacy means the state of an individual or group being able to seclude themselves or their information. 'Security' is a broad term that includes secure access and confidentiality. Data 'confidentiality' is not just a security issue, but also juristic concerns (McGuire et al., 2008). Health information is considered the most confidential of all types of personal information. Protecting individual privacy and maintaining confidentiality are attracting increasing attention from healthcare organisations and governments (Churches, 2003). Therefore, data security is broken down into 'confidentiality' and 'secure access'. This suggestion is supported by the review of ISO 18308 detailed in Table 3-1.

Plausibility

This dimension of data quality is mentioned only by Weiskopf and Weng (2013) and, despite the fact that it was considered a separate dimension, they did not consider it fundamental and it serves as a proxy for the other dimensions. Moreover, this dimension is not commonly used in the literature on data quality, so the dimension was disregarded.

Subsequently, Health Informatics requirements for an electronic health record architecture, ISO 18308 (British Standards Institute, 2011) was reviewed to identify its requirements against the dimensions. This review resulted in Table 3-1, confirming that these dimensions assess the data populating EHR systems.

Table 3-1. Data quality dimensions and corresponding ISO 18308 codes

DIMENSIONS	CORRESPONDING ISO18308 CODES
Accuracy	QTY4, HRP4 and SAC3
Completeness	VER11
Consistency	HSO5 and HSO6
Timeliness	RET2 and HRP4
Relevance	CPO2, HRP1, SBJ1, SBJ2, IAA5 and IAA6
Usability	HSO2, HSO3, LIN6, RET8, SAC2 and CPO5
Provenance	HRP1-6
Interpretability	CPO6, STR11 and TXT2-7
Secure Access	IAA1-15, ACT1-3 and CON1-6
Privacy	
Confidentiality	

Other dimensions (interpretability, provenance)

Although interpretability is not added to the frameworks developed by the Canadian Institute for Health Information (2009) or Kerr et al. (2007) as a dimension of data quality in the health context, they consider it as an essential criterion to measure their usability dimension. Also, it has been recognised by many data quality methodologies and frameworks (Wang 1998; Jarke et al., 1998; Pipino et al., 2002; Lee et al., 2002; Su & Jin, 2006) as an important facet of data quality. In this research, interpretability is a used as a dimension due to the fact that uninterpretable data is one of the challenges of using EHRs (Bayley et al., 2013).

With regard to provenance, this dimension facilitates showing the source and the derivation of EHR data. Orfanidis et al. (2004) point out the importance of provenance in data quality assurance processes for EHR systems. They include it within the set of data quality requirements specific to EHRs in their study. ISO 18308 (British Standards Institute, 2011) emphasises the essence of health record provenance, and reveals that EHRs are supposed to 'represent health record entries in a way that provides the accurate chronology of their authorship and availability within the EHR'. It also stresses the need for the unaltered persistence of an original data entry, and keeping track and noting any subsequent change to or deletion of that data entry. Moreover, Moreau et al. (2008) illustrate how provenance is essential in the health field as it allows powerful queries that help not only medical profession but also regulators and the families of the patient.

3.1.2 Data quality categories

The category is the overall label to which a group of dimensions belong. It is a useful way of ensuring complete coverage of concept of data quality (Lee et al., 2002). The most common categories for data quality dimensions are the terms intrinsic, contextual, representational and accessibility coined by Wang & Strong (1996). However, this classification and grouping caused some overlap as some dimensions can fall under more than a category as detailed in Section 2.2.4.

For data to be fit for use, data or information products must possess the three attributes of quality, namely utility, objectivity and integrity (Tupek, 2006). This implies that any defined dimension should belong to only one of these attributes.

- Utility: '*refers to the usefulness of the information for its intended users*';
- Objectivity: '*refers to whether information is accurate, reliable, and unbiased, and is presented in an accurate, clear, and unbiased manner*';
- Integrity*: 'refers to the security or protection of information from unauthorized access or revision'.*

Thus, the characteristics of high-quality data as discussed in Section 3.1.1 are mapped into the three categories: objectivity, utility, and integrity. As can be seen from Figure 3.4, there are 11 data quality dimensions in a framework of three categories.

3.1.3 Proposed data quality dimensions

This Section outlines the proposed framework and its 11 dimensions, mapped to the relevant categories. It also provides definitions for the dimensions of the framework.

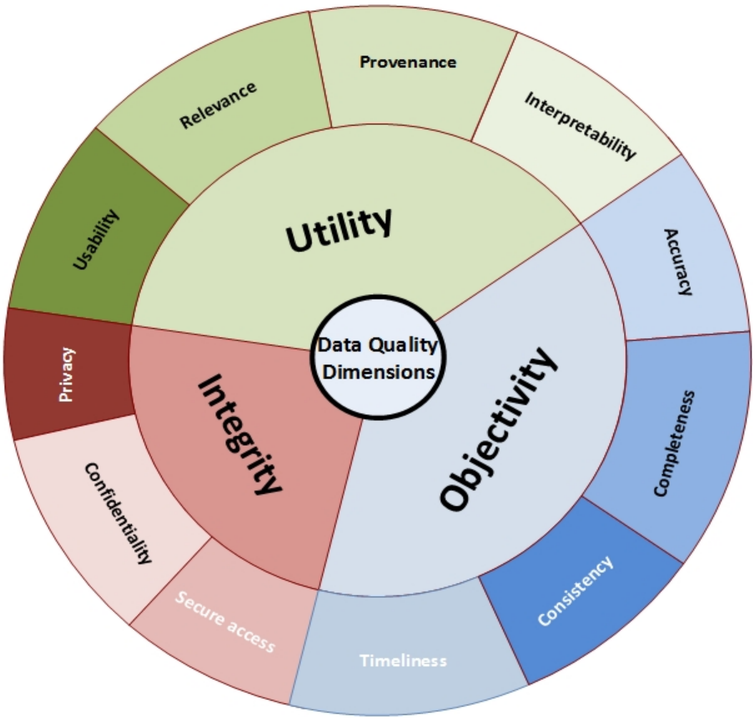


Figure 3.4. The framework of data quality in EHR

3.1.3.1 Objectivity

Objectivity is one of the three attributes that any data must possess to achieve high quality. Its dimensions are detailed below:

Accuracy: The extent to which registered data conforms to its actual value.

Completeness: The state in which information is not missing and is sufficient for the task. Linkages between data promote the existence of further data.

Consistency: Representation of data values remains the same in multiple data items in multiple locations.

Timeliness: The state in which data is up to date and is availability on time.

3.1.3.2 Utility

Utility is the second characteristic attributed to data that are considered to be of high quality. It concerns the usefulness of the data for intended users. Because of this communication, EHR systems have multiple data items in multiple locations.

Relevance: The extent to which information is appropriate and useful for the intended task.

Usability: The ease with which data can be accessed, used, updated, understood, maintained and managed.

Provenance: The source of data, shown and linked to metadata about data.

Interpretability: The degree to which data can be understood.

3.1.3.3 Integrity

Security prevents personal data from being corrupted and controls access to ensure privacy and confidentiality. Its components are as follows:

Secure Access: Personal data being protected against unauthorised access.

Confidentiality: The state of information being secret or accessibly restricted under a set of rules that limits the access.

Privacy: The state of an individual or group being able to seclude themselves or their information.

3.2 Data Quality Assessment

Quality assessment is a crucial phase in data quality assurance. It helps any organisation to assess the status of its organisational data and to plan improvement activities. Two kinds of assessment, subjective and objective, need to be involved. Subjective assessment evaluates the quality of data from the perspective of stakeholders; that is, data collectors, data custodians and data consumers (Lee et al., 2002). It captures their opinions on whether it is fit for use. Assessment tools for this kind of assessment could be questionnaires, interviews or surveys.

Objective assessment is cost effective, as the process can be automated and have no need for intervention. Lack of bias is another feature of objective assessment, as opinions and attitudes are not involved. This kind of assessment could be either task-independent or task-dependant (Pipino et al., 2002). Task-independent measures can measure the quality of dataset without contextual knowledge of the application, while task-dependent measures can measure any violations to the business rules and regulations of an organisation/government department and the database constraints imposed by administrators (Cappiello et al., 2003).

For the sake of accuracy and automation, objective assessment is recommended. However, some dimensions can only be assessed subjectively, such as relevance and interpretability. Ge and Helfert (2008) provide a model in which they group the data quality dimensions provided (Wang & Strong, 1996) into two categories, based on assessment type. Accordingly, accuracy, consistency, completeness and timeliness are classified as objective assessments, while the rest are subjective assessments.

3.2.1 Measures for objective assessment

As discussed in Section 2.2.9, many studies have been undertaken exclusively to highlight problems that compromise the data quality and, subsequently, have led to different taxonomies of data quality problems. The existing taxonomies help organisations to consider all problems, and present a comprehensive set of data quality problems. Rationally, data quality problems should fall under the dimensions proposed in the literature to facilitate the assessment process as a dimension is a facet of data quality. However, there is no taxonomy of 'dirty data' that groups quality problems according to their relation to dimension measurement. In this Section, a dimension-oriented taxonomy of data quality problems is proposed that would help organisations to assess each dimension of quality and prioritise them.

The taxonomy shown in Figure 3.5, proposed by Rahm and Do (2000), was adopted as an initial collection of data quality problems. Subsequently, other types of 'dirty data' that are found in other studies were thoroughly analysed and filtered to identify those not included in that initial collection.

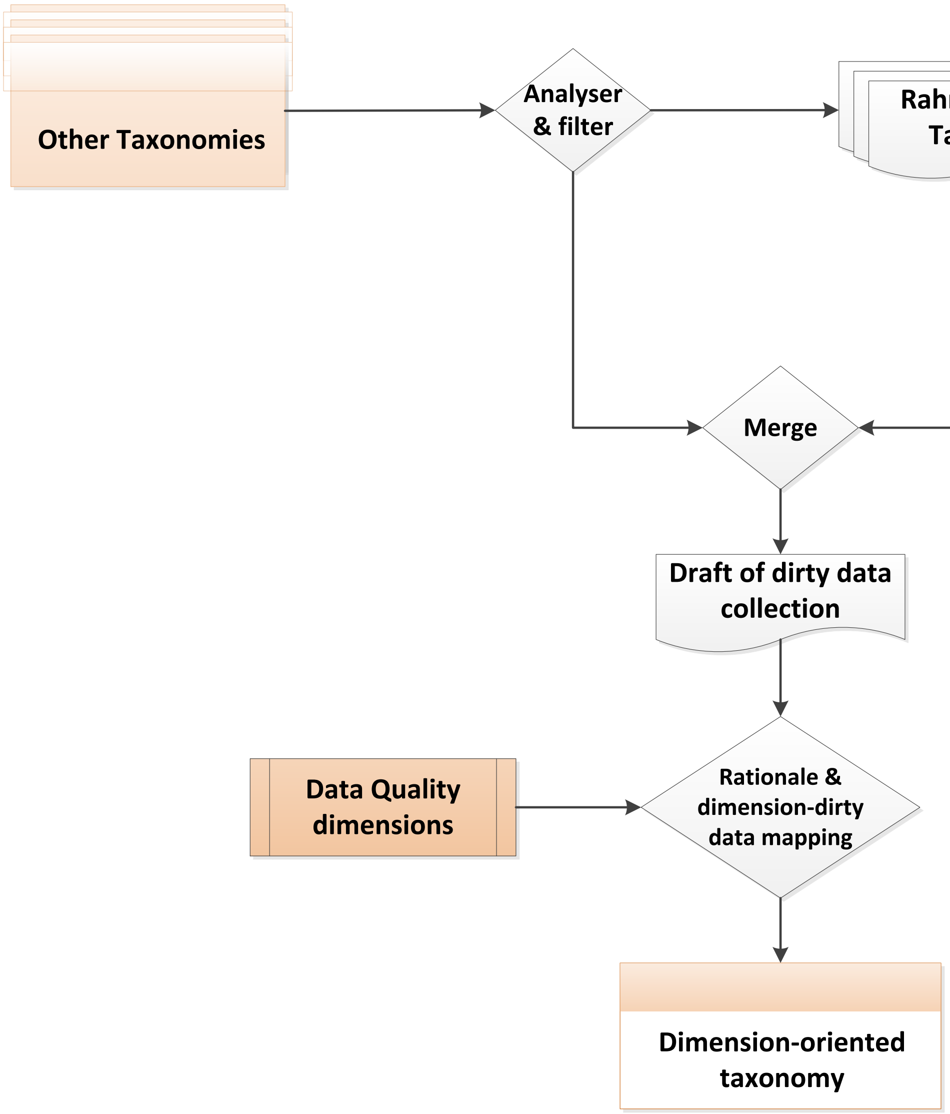

Other Taxonomies
Analyser & filter
Rahm and Do's Taxonomy
Merge
Draft of dirty data collection
Data Quality dimensions
Rationale & dimension-dirty data mapping
Dimension-oriented taxonomy

Figure 3.5: Process of developing a dimension-oriented taxonomy

Therefore, we ended up with a draft of the types of 'dirty data' that cover all aspects of errors mentioned earlier in the literature. The last step was a rationale stage in which each type of 'dirty data 'was examined against data quality aspects (dimensions). Consequently, each dimension has respective types. Table 3-2 shows the dimension-oriented taxonomy of 'dirty data'.

Table 3-2: Dimension-oriented taxonomy of 'dirty data'

NO	Data Quality Problems	Dimension
D1	Illegal values due to invalid domain range	ACCURACY
D2	Misspellings	
D3	Misfielded values	
D4	Embedded values	
D5	Word transposition	
D6	Wrong reference	
D7	Erroneous entry	
D8	Violated attribute dependencies	CONSISTENCY
D9	Uniqueness violation	
D10	Naming conflicts in multi-source	
D11	Structural conflicts in multi-source	
D12	Wrong categorical data	
D13	Relational integrity violation	
D14	Violated attribute dependencies	
D15	Duplicated records in single/multi data source(s)	
D16	Contradicting records in single/multi source(s)	
D17	Inconsistent spatial data	
D18	Different measure units in single/multi source(s)	
D19	Syntax inconsistency	
D20	Missing data where Null-not-allowed constraint enforced	COMPLETENESS
D21	Missing data where Null-not-allowed constraint not enforced	
D22	Missing record	
D23	Ambiguous data due to incomplete context	
D24	Semi-empty tuple	
D25	Outdated temporal value	TIMELINESS
D26	Outdated reference	
D27	Different representations due to use of abbreviation and cryptic values	INTERPRETABILITY
D28	Different representations due to use of alias/nickname	
D29	Different representations due to use of encoding format	
D30	Different representations due to use of special characters	

3.2.2 Measures for subjective assessment

There are many data quality methodologies that select subjective measures in the assessment phase. AIMQ methodology (Lee et al., 2002) is widely recognised in the data quality context. It comprises three components: the PSP/IQ model, the IQ instrument and gap analysis techniques. Each characteristic includes several information quality criteria. This methodology is based on subjective measures and uses a questionnaire designed to capture users' opinions for each criterion on a scale of 0 to 10.

For the integrity-related dimensions, 'relevance dimension' and 'usability dimension' of the proposed framework, the researcher adapted quality measures and the methodologies from AIMQ with some changes and checked its compliance with the technical definitions of the respective dimensions. The scale

0 to 10 is adopted to capture as well the opinions and experience of data consumers. The measures of usability are adapted from CIHI, see Section 2.3.3.

Table 3-3: Subjective measures for subjective dimensions

NO	Data Quality Measures	Dimension
1	Is this information easily accessible	Usability
2	Is this information easily retrievable	
3	Is this information not usable	
4	Is this information promptly accessible when needed	
5	Is this information easily updatable	
6	Is this information easily understood	
7	Is this information usable	
8	Is the availability of information for the patient treat adequate	
9	Is this information easily manipulated	
10	Is this information relevant to the task at hand	Relevance
11	Is this information useful to the task at hand	
12	Is this information applicable to the task at hand	
13	Is this information appropriate for the task at hand	
14	Is this information irrelevant to the task at hand	
15	Is the origin of this information clearly exist	Provenance
16	Is this information owned by known subject	
17	Is the creation date of this information shown	
18	Is the update history of this information exist	
19	Is the access to this information sufficiently restricted	Secure Access
20	Does the access to this information require authentication process	
21	Is the owner/creator of this information authorised	
22	Is unauthorised access to this information sufficiently prevented	
23	Is the confidentiality of this information achieved	Confidentiality
24	Is this information only accessible by authorised people	
25	Is the access to this information granted only to persons who should see it	
26	Is this information vulnerable	
27	Is the sensitivity of this information clearly declared	
28	Could this piece of information be disseminated without permission	
29	In case of sharing, is there a clear consent for this information to be shared	
30	Is personally identifiable private information is being appropriately safeguarded	Privacy
31	Is this personal identifiable information compliant with privacy policy	
32	Is the policy regulation obeyed in this piece of information	
33	Is privacy policy violated in this piece of information	
34	Is the access to this personal identifiable information granted only to persons who should see it	
35	In case of an individual being identified, is there a clear consent from this individual to be identified	

With regard to provenance measures, the provenance requirements published within ‘Health information – Requirements for an electronic health record architecture, ISO 18308:2011’ were analysed to facilitate the development of measures for the dimension of ‘provenance’. The requirements address four objectives: the origin of the information; the ownership of the information; the creation date and time of information; and the update history of the information. These objectives were considered as measures of 'provenance'. Table 3-3 shows the subjective measures generated for dimensions that are assessed subjectively.

3.2.3 Measures and aggregation methods

In objective assessment, multiple measurable items are provided for all objective dimensions. The values derived from these measures are aggregated and used to calculate the quality score for the relevant dimension.

The measures associated with accuracy, completeness, consistency, timeliness and interpretability are presented in Table 3-2. Naumann et al. (1999) provide an operational definition for accuracy, which they characterise as 'the percentage of objects without data errors such as misspellings, out-of-range values, etc.' Boolean measure is the type of measure used to measure the accuracy of individual data items. That is, if a data item is free of a data quality problem (e.g. misspellings), the Boolean value is (1), otherwise it is (0). This is applied to all objective dimensions. Accordingly, the measure is:

$$\text{Dimension quality} = avg\left(\sum_{i=1}^{n} \frac{no.of\ data\ units\ free\ of\ M_i}{total\ no.of\ data\ units\ assessed}\right)$$

where M_i is the ith measure within a dimension, and n is the number of measures associated with the dimension.

It is worth mentioning that a data unit can be an attribute of a tuple (a cell) or a tuple, depending on the data quality problem (measure).

By contrast, in subjective assessment user evaluation of the data are used. A value expressed in the range of 0 to 10 has been employed and has proved its usefulness in data quality measures (Lee et al., 2002). Table 3-3 displays the measures (data quality items) associated with each subjective dimension. To capture the user or consumer evaluation, a questionnaire featured 35 measures for 'usability', 'relevance', 'provenance', 'secure access', 'confidentiality' and 'privacy' on a 11-point Likert scale bounded by 'not at all' (=0) and 'completely' (=10).

In the aggregation method, subjective assessments tend to involve multiple users from different management levels in order to produce rigid results. Thus, the measure for the subjective assessment is:

$$\text{Dimension quality} = avg(\sum_{i}^{n} avg\left(\sum_{j=1}^{p} S_{ij}\right)$$

where n is the number of measures associated with a dimension, and p is the number of participants involved in the assessment process. S_{ij} is the value of the assessment of the ith measure and jth participant.

3.3 Chapter summary

The main purpose of this chapter was to review the data quality dimensions and their associated measures in EHR and healthcare field. This led to identifying potential dimensions and measures. The methodology of developing the proposed framework and respective measures were discussed in Sections 3.1 and 3.2. A confirmatory study (explained in the next chapter) was conducted to confirm the developed framework of data quality dimensions.

Chapter 4 Research Methodology

In previous chapters, the focus was on literature relating to the aspects of data quality in EHRs, measures for objective and subjective assessment, and the proposed framework and its associated measures. This chapter discusses the research methods used in confirming the dimensions of data quality in EHR systems in Saudi Arabia and developing the measures associated with these dimensions to answer the research questions.

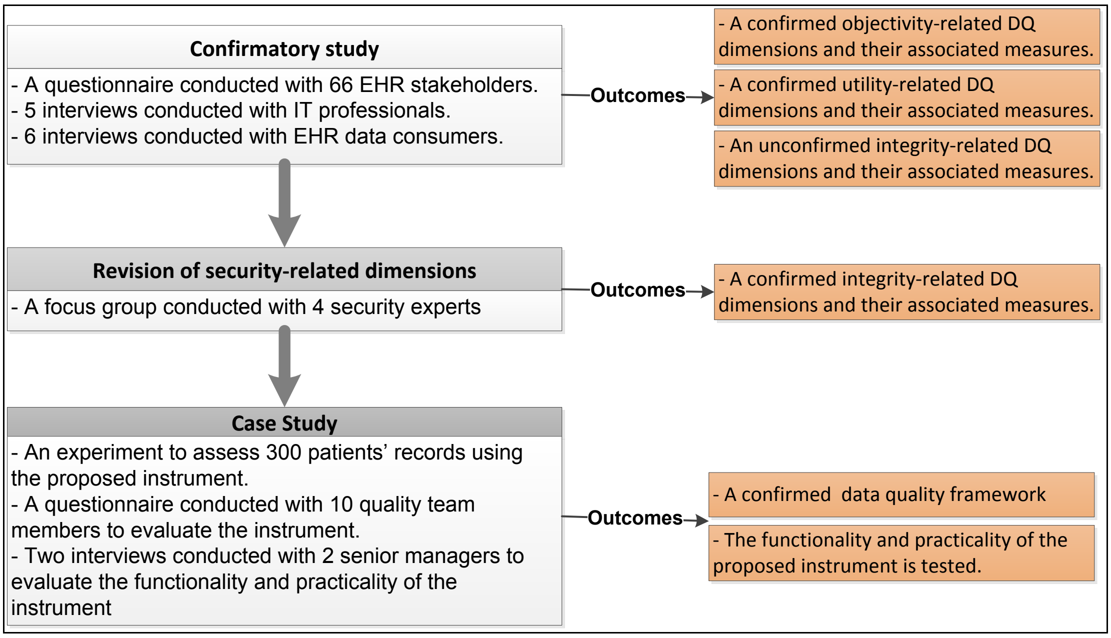

Confirmatory study
- A questionnaire conducted with 66 EHR stakeholders.
- 5 interviews conducted with IT professionals.
- 6 interviews conducted with EHR data consumers.
Outcomes
- A confirmed objectivity-related DQ dimensions and their associated measures.
- A confirmed utility-related DQ dimensions and their associated measures.
- An unconfirmed integrity-related DQ dimensions and their associated measures.
Revision of security-related dimensions
- A focus group conducted with 4 security experts
Outcomes
- A confirmed integrity-related DQ dimensions and their associated measures.
Case Study
- An experiment to assess 300 patients' records using the proposed instrument.
- A questionnaire conducted with 10 quality team members to evaluate the instrument.
- Two interviews conducted with 2 senior managers to evaluate the functionality and practicality of the instrument
Outcomes
- A confirmed data quality framework
- The functionality and practicality of the proposed instrument is tested.

Figure 4.1: conducted research methods and stages

The chapter contains several Sections as outlined in Figure 4.1. Section 4.1 gives an overview of the research questions, sub-research questions and the methods used for answering these questions. A brief explanation of the methods used in this research is presented in Section 4.2. Section 4.3 consists of several Sections; each Section gives details on how a research method used was designed. Section 4.3.1 presents the interview method including the selection of the experts and EHR consumers, the data collection process and the qualitative data analysis. Section 4.3.2 presents the questionnaire design used to confirm the proposed framework. It explains the data collection process, quantitative data analysis and the sample size determination. The design of a focus group including the selection of the quality assessment team members, data collection process and data analysis process, is presented in Section 4.3.3. Section 4.3.4 presents the design of the case study. It details the context of the study and participants involved in the assessment of EHR data of the selected hospital. It also explains how the survey used to evaluate the practicality of the proposed instrument was developed.

4.1 Overview of Research Questions

The research methods were conducted to answers the three research questions, RQ1, RQ2 (three sub-research questions) and RQ3. Each research question and related sub-research questions, and the methods for answering the sub-research questions, are shown in Table 4-1.

Table 4-1: Research questions, sub-research questions and methods for answering

Research question	Sub-research question	Method for answering sub-research question
RQ1: What data quality determinants are important for EHR stakeholders perceived data quality?	-	• Review of frameworks. (Chapter 1) • A questionnaire with 66 EHR people. (Section 4.3.2) • Semi-structured interviews with 5 IT professionals. (Section 4.3.1) • Semi-structured interviews with 6 EHR data consumers. (Section 4.3.1)
RQ2: Based on the proposed framework, what is the appropriate instrument with which to measure quality of data in EHRs?	RQ2.1: What are the measuring items for objective data quality assessment?	• Semi-structured interviews with 5 IT professionals. (Section 4.3.1) • Semi-structured interviews with 6 EHR data consumers. (Section 4.3.1)
	RQ2.2: What are the measuring items for subjective data quality assessment?	• Semi-structured interviews with 5 IT professionals. (Section 4.3.1) • Semi-structured interviews with 6 EHR data consumers. (Section 4.3.1) • A focus group with 4 security experts. (Section 4.3.3)
	RQ2.3: How well is the functionality and practicality of the proposed instrument?	• A case study (Section 4.3.4) • An evaluation survey conducted on the quality team members. (Section 4.3.4) • Two semi-structured interviews with senior managers. (Section 4.3.4)
RQ3: what are the severity factors that make data quality problems more severe?	-	• Semi-structure interviews with 5 IT professionals. (Section 4.3.1) • Semi-structure interviews with 6 EHR data consumers. (Section 4.3.1)

4.2 Research Methods

This Section provides a brief description about the many research methods used in this study to answer the research questions, including a literature review, questionnaires and semi-structured interviews. It explains the triangulation techniques used to confirm the proposed data quality framework. The questionnaire was developed using a five-point Likert scale. The questionnaire data were collected from individuals who work for health provider organisations including Health Ministry officials, directors, medical consultants, GPs, nurses and pharmacists. Semi-structured interviews were conducted with IT experts, medical directors, consultants and a nurse in King Abdulaziz Medical City.

4.2.1 Qualitative research

Qualitative research is primarily exploratory and is used to uncover the perceptions of the target audience with reference to specific topics or issues. It helps researchers understand phenomena by providing insights into problems or producing hypotheses for potential quantitative research (Berg & Lune, 2012). Its participants are usually fewer in number than those in quantitative research (Hancock et al., 1998).

An interview is considered the most common method through which qualitative data are collected, and the semi-structured interview is one the four types that mix structured and unstructured questions (Drever, 2003; Rogers et al., 2011).

4.2.2 Quantitative research

Quantitative methodology is used to quantify numerical data, or data that can be transformed into usable statistics, by surveying a large number of participants. It quantifies the attitudes, opinions, behaviours or other defined variables. Subsequently, the data gathered is analysed by statistical techniques (Berg & Lune, 2012). The result, therefore, is more generalizable to a population (Thomas, 2003). A Likert scale is the most commonly used scale in quantitative research, and is used to capture the opinions of a subject (Saunders et al., 2009).

4.2.3 Triangulation

Triangulation techniques are used to validate and confirm the data quality of the proposed framework. The term refers to the method of combining methodologies to facilitate the validation of research findings (Jick, 1979). For the confirmation of the data quality framework, data triangulation includes a review of relevant frameworks in literature, user validation (questionnaires) and interviews with experts and data consumers. For the validation of quality items making up each dimension, data triangulation includes the literature review, expert validation and data consumer validation.

4.2.4 Case study

A case study is a technique that provides timely insights and deeply focuses on a range of contexts, organisation or people (Gray, 2009). It is also a distinctive method that can serve evaluation needs directly through assessing outcomes and testing hypothesis (Yin, 1992). A tool evaluation that is based on a case study is characterised as a focused, in-depth analysis and synthesis of a particular instrument or program. This type of method is highly appropriate in program (instrument) evaluation (Stufflebeam, 2001).

4.3 Design of the Research Methods Used in this research

This Section describes the research methodologies utilised in this study to facilitate the confirmation of the findings and to discover possible potential dimensions for data quality assessment. The questionnaire is a type of quantitative research approach used to make a generalisation from a sample to the whole population. It was developed to collect data from people working for health providers' organisations to confirm the proposed data quality dimensions. The interview was designed using qualitative research techniques to confirm the proposed dimensions and their quality items and explore more potential dimensions and their items. Interviews were carried out with both experts and data consumers. A revision of integrity-related dimensions and associated measures is an emerging research step due to no conclusions reached for these dimensions. Thus, a focus group was utilised to review the intended dimensions and their associated measures. Lastly a case study was conducted to evaluate the functionality and practicality of the proposed instrument. Within the case study, 10 quality team members were surveyed and two interviews were carried out with two seniors.

4.3.1 Interview design

Semi-structured interviews were utilised to collect data from two groups. This kind of interview was selected due to its advantage of gathering statements regarding the individuals' attitudes and exploring in-depth their experience (Drever, 2003). It was used to capture their experience regarding:

- The aspects of data quality
- The problems associated with data
- How to assess the quality of data.

The following subsections briefly describe how the semi-structured interviews were designed.

4.3.1.1 Selection of experts

This study was conducted at National Guard Health Affairs (NGHA) in Saudi Arabia in February 2013. NGHA is one of the leading health organisations providing healthcare to National Guard employees and their dependants. It was chosen first as it has under its umbrella four hospitals and 60 primary and secondary health centres across Saudi Arabia. Secondly, it received the Middle East Excellence Award in EHR in 2010[8].

The interviews were conducted with two groups. The first was of IT experts with responsibility for implementing and maintaining EHR systems, comprising five IT professionals with various responsibilities belonging to the Information Services and Informatics Division (ISID) and the Clinical Information Management Systems (CIMS). Table 4-2 shows snapshot of the IT experts interviewed in this study.

Table 4-2: Selected expert Interviewees

Participant	**Position**	**Experience (years)**	**Justification**
Senior DBA1	Database administrator	+15	Direct involvement with quality problems in databases
Senior DBA2	Database administrator	+10	Direct involvement with quality problems in databases
Team Leader1	Physician team leader	+5	Link between medical staff and IT support
Team Leader2	Physician team leader	+5	Link between medical staff and IT support
Analyst	Application analyst	+5	Direct involvement with application processes and duty of application enhancement

The other group consisted of data consumers. They were all selected from King Abdulaziz Medical City (KAMC). The reason for choosing staff from here was that it was accredited under Joint Commission International standards (JCI) in 2006 with excellent performance. The other reason was that its staff are highly qualified and well-trained, and some of them hold academic positions at King Saud bin Abdulaziz University for Health Sciences. Table 4-3 shows a snapshot of the data consumers selected for interview.

Table 4-3: Selected data consumers for interview

Participa nt	**Position**	**Experience** (years)	**Justification**
CONSULT1	Paediatric consultant	+15	Main decision maker in patients' medical care
CONSULT2	Radiology consultant	+10	Data generator and decision maker in patients' medical care
CONSULT3	Paediatric consultant and Assistant professor	+15	Main decision maker in patients' medical care
Med Director	Medical director	+10	Decision maker in policy and work regulations
CONSULT4	Emergency consultant	+10	Decision maker in a very critical department
Nurse	Nurse	+13	Data generator due to direct involvement in patient medical care

[8] http://www.ngha.med.sa/English/Pages/ArabHealthAward.aspx

4.3.1.2 Data collection process and instrument

The interviews with the two groups were scheduled over two weeks. After introducing the research background and purpose, a consent form was given to the interviewees to sign to assure their agreement to participate in the study. The interviewees were asked for 11 dimensions. Each dimension consisted of nine questions. The interview featured confirmatory and exploratory questions about the dimensions and their quality items making up the metric (see Appendix C). Livescribe[9] pen was used as a tool for recording the interviews.

4.3.1.3 Qualitative data analysis

Thematic analysis was used to analyse, identify and report the themes within raw data. The themes reflect patterns exist within the collected data, and the patterns describe the phenomenon. Therefore, it is a method of organising and describing a corpus in a way that help researchers capture important things to describe their research questions (Aronson, 1994; Braun & Clarke, 2006).

As the interview questions revolve around data quality dimensions and their quality items. Therefore, themes and sub-themes were dimensions, and the sub-themes address any related issue. To facilitate the qualitative data analysis, Nvivo 10 software was utilised to theme raw data. Each dimension was given a node, each node has its characteristics and its quality items clustered into 'confirmed', 'irrelevant', 'additional' and 'overlapping'. The next step was to code and assign data from the transcript to related nodes.

Besides, there was a node for the severity factors to which every statement concerning about severity factors link. That is, any statement that describes a state where the data quality problems become severer would be coded and assigned into the node of severity factors.

4.3.1.4 Ethical Approval

The quantitative and qualitative methodologies conducted in this study were approved by the Ethical Committee of the School of Electronic and Computer Science at the University of Southampton. Ethical approval was granted under reference number **ERGO/FPSE/9382**.

4.3.2 Questionnaire design

A survey is used to collect information to capture knowledge, attitudes and behaviours. Questionnaires are a data collection tool in which participants are requested to answer various predetermined questions. There are two types of questionnaire: self-administered and interview-administered. A self-administered survey is one in which the respondents take responsibility for reading and

[9] http://www.livescribe.com/uk/

answering the questions, while the communication medium used in interview-administered surveys is either a personal or a telephone interview (Zikmund et al., 2012).

4.3.2.1 Data collection process and instrument

The questionnaire featured 11 identified determinants on a five-point Likert scale bounded by 'very important' (=1) and 'not at all important' (=5). The communication medium for data collection was self-administered via an online software tool called SurveyGizmo.[10] Subsequently, the questionnaire was distributed to Saudi National Health Services (NHS) staff to capture their attitudes and opinions towards the identified dimensions and its ability to assess the quality of EHR data properly. The questionnaire (see Appendix A) was divided into three Sections. The first was to collect demographic data about the participant. The second Section was to measure the NHS staff members' opinions on the importance of the proposed items. The last Section was developed to check if there is ambiguity in this survey, and whether any data quality dimension needed to be incorporated.

4.3.2.2 Quantitative data analysis

One-sample T test is a statistical procedure used to test the mean value of a distribution. For this study, to accept a factor as a reliable dimension, the mean value of a representative question of this factor needs to be significantly lower than 2.5. The rationale behind choosing 2.5 is that this number falls midway between 'important' and 'neutral' on the five-point Likert scale.

To conduct the hypothesis test for this statistical test, a confidence level of 95% was used, and the error rate accepted by the researcher α (alpha) was set at 0.05. So the null hypothesis and alternative hypothesis were as follows:

(H_0: the mean value of a factor is higher than 2.5)

(H_1: the mean value of a factor is equal or lower than 2.5).

The factor is accepted as a sound dimension for this study if the null hypothesis was rejected ($\rho<0.05$).

4.3.2.3 Determination of minimum sample size

Statistical power analysis assisted the researchers in calculating the minimum sample size for this study. We needed to consider an error of the first kind (Type I error) and an error of the second kind (Type II error). Type I errors (α)

[10] http://www.surveygizmo.com/

occur when the null hypothesis is true, but wrongly rejected, while Type II errors (β) occur when the null hypothesis is false, but not rejected.

G*Power software was used to facilitate acquisition of the minimum sample size to reject H_0. Table 4-4 shows the sample size we needed for this study.

Table 4-4: Sample size needed for the study

Tail(s)	2
Effect size	0.5
α err prob.	0.05
Power (1- β err prob.)	0.95
Minimum sample size	54

4.3.2.4 Ethical Approval

The quantitative and qualitative methodologies conducted in this study were approved by the Ethical Committee of the School of Electronic and Computer Science at the University of Southampton. Ethical approval was granted under reference number **ERGO/FPSE/6327**.

4.3.3 Focus group design

A 'focus group' was the technique used in this study to collect and analyse data from a group of security experts. A focus group is defined by Carey (1994) as a session 'using a semi structured group session, moderated by a group leader, held in an informal setting, with the purpose of collecting information on a designated topic'. The purpose of the focus group is to assure and examine the dimensionality of security-related aspects of data quality, as well as the reliability of their associated measures.

Most authors recognise the main advantage of focus groups over other types of interviews as being the use of interaction to generate data (Kahn et al., 1991; Kitzinger, 1996; Morgan, 1996; McLafferty, 2004). It is considered as a potential technique for developing a new instrument (Gray-Vickrey, 1993; McKinley et al., 1997). Another characteristic of a group interview is that it facilitates the validity enhancement of instruments (Powell et al., 1996).

The researcher chose this focus group type of interview over individual interviews for many reasons, including that the participants would have the opportunity to exchange anecdotes, and to challenge and comment upon the ideas and opinions of others (Kitzinger, 1995). Besides, discrepancies tend to be more associated with individual interviews than with focus groups (Morgan, 1996), so adopting a focus group as data collection tool could lead to a greater degree of validity and consensus on the items, as is needed for this type of research (Rubin & Rubin, 2011). Moreover, the interactive discussion of a focus group would overcome the ambiguity issue that the researcher encountered detailed in Sections 6.1 and 6.2 (Hennink et al., 2010).

4.3.3.1 Selection of experts

To enrich the discussions regarding dimensionality of the security facets as well as the measuring items of those aspects, the technical experiences and theoretical grounds of security were considered in the selection of the experts. This was to give an appropriate balance between theory and practice in the discussion.

Sample size is an endless issue, as there is little agreement on what is most appropriate. Morgan (1996) believed that small groups, usually from four to six participants, are more likely to interact better and be easy to manage. Many studies (Kitzinger, 1996; Twinn, 1998; Saunders et al., 2009) have proposed that four is the minimum number of participants in a focus group.

Thus, the focus group was carried out with four security experts in health field; two of the experts have a strong technical background, while the other two have good experience and knowledge in security-related ISO standards and models.

Table 4-5: Participants in the focus group

Participant	Experience(years)	Institution
Expert A	+15	Hospital
Expert B	+10	Health Authority
Expert C	+5	Ministry of Health
Expert D	+10	Saudi Food and Drug Authority

4.3.3.2 Data collection process

The focus group was conducted at the Saudi Food and Drug Authority, and the duration of the meeting was two hours. Prior to the discussion, the participants were introduced to the research background and the purpose of the meeting. Subsequently, the participants were given a consent form to sign to indicate that they understood the policies of the interview, and agreed to participate in the study.

Discussions during the group interview focused on obtaining constructive feedback first on the security dimensions, their definitions and implications, and second to the measuring items associated with the predefined dimensions. The first area covered the three dimensions – secure access, confidentiality and privacy – as detailed in Table 4-6, while the other exposed the integrity dimensions-related measuring items detailed in Table 4-7.

Table 4-6: Integrity-related dimensions and their definitions

Dimension	Definition
Secure access	Data being protected against unauthorised access
Confidentiality	The state of information being secret or accessibly restricted under a set of rules that limits the access
Privacy	The right and desire of a person to control the disclosure of personal health information.

Table 4-7: Measuring items associated with integrity-related dimensions

Secure Access: Data being protected against unauthorised access

- Is the access to this information sufficiently restricted
- Does the access to this information require authentication process
- Is the owner/creator of this information authorised
- In case of system failure, is data safely recoverable
- In case remote access, is remote access policy applied
- Is this information vulnerable

Confidentiality: The state of information being secret or accessibly restricted under a set of rules that limits the access

- Is the confidentiality of this information achieved
- Is this information only accessible by authorised people
- Is the sensitivity of this information clearly declared
- Could this piece of information be disseminated without permission
- In case of sharing, is there a clear consent for this information to be shared

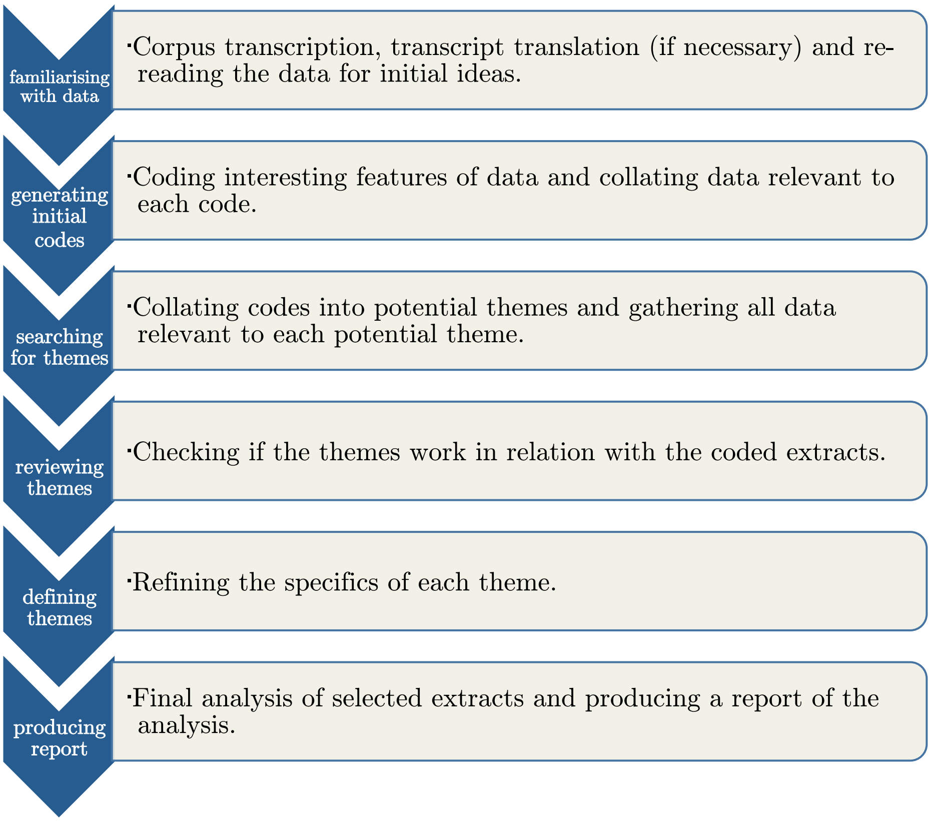
Privacy: The right and desire of a person to control the disclosure of personal health information.

- Is personally identifiable private information being appropriately safeguarded
- Is this personally identifiable information compliant with privacy policy
- Is the access to this personally identifiable information granted only to persons who should see it
- In case of an individual being identified, is there a clear consent from this individual to be identified

Despite their fluency and proficiency in English, the communication language used in the discussion was Arabic as it was their preferences. A video recorder was used, being the recommended data collection strategy to facilitate the process of gathering non-verbal and verbal data (Polgar & Thomas, 2013).

4.3.3.3 Data analysis procedure

Thematic analysis is a foundational method for qualitative analysis (Braun & Clarke, 2006). It is a technique used to identify and classify the qualitative data collected from the focus group into themes and categories. The themes and categories reflect patterns within the collected data, and the patterns describe the phenomena. Therefore, it is a method of organising and describing a corpus in a way that helps researchers to capture important things to describe their research questions (Aronson, 1994; Braun & Clarke, 2006). The stages of the thematic analysis are illustrated in Figure 4.2.



·Corpus transcription, transcript translation (if necessary) and re-reading the data for initial ideas.

·Coding interesting features of data and collating data relevant to each code.

·Collating codes into potential themes and gathering all data relevant to each potential theme.

·Checking if the themes work in relation with the coded extracts.

·Refining the specifics of each theme.

·Final analysis of selected extracts and producing a report of the analysis.

Figure 4.2: Phases of thematic analysis summarised from (Braun & Clarke, 2006)

A theme represents something important about qualitative data, in relation to the research question. Themes within qualitative data can be identified data in either inductive or theoretical thematic analysis. In the inductive approach, the themes identified bear little relation to the questions asked during interviews, as they are data-driven, while in theoretical thematic analysis a theory-driven or analyst-driven analysis provides a more detailed analysis of some aspects of data, and not of the data overall (Braun & Clarke, 2006). As the aim of this study was primarily to confirm the identified security aspects of data quality as well as their associated measures, theory-driven analysis was selected to analyse our data, and would use existing frames to help the researcher to capture the participants' opinions on the dimensions and their measures.

4.3.3.4 Data preparation for analysis

Data preparation involves corpus transcription, transcript translation into English and setting themes based on the main goals of the study. NVivo10 software was utilised to facilitate the qualitative data analysis. As the confirmatory study revolves around three integrity-related dimensions and their measures, three nodes were set; a node for each dimension. Also, sub-nodes were developed for associated measures. Sub-nodes were introduced under each node of integrity-related dimension in order to avoid failing to identify any potential measures for the integrity-related dimensions. A 'potential security aspect' code and associated measures were introduced to avoid failing to identify other potential security facets.

4.3.3.5 Data analysis process

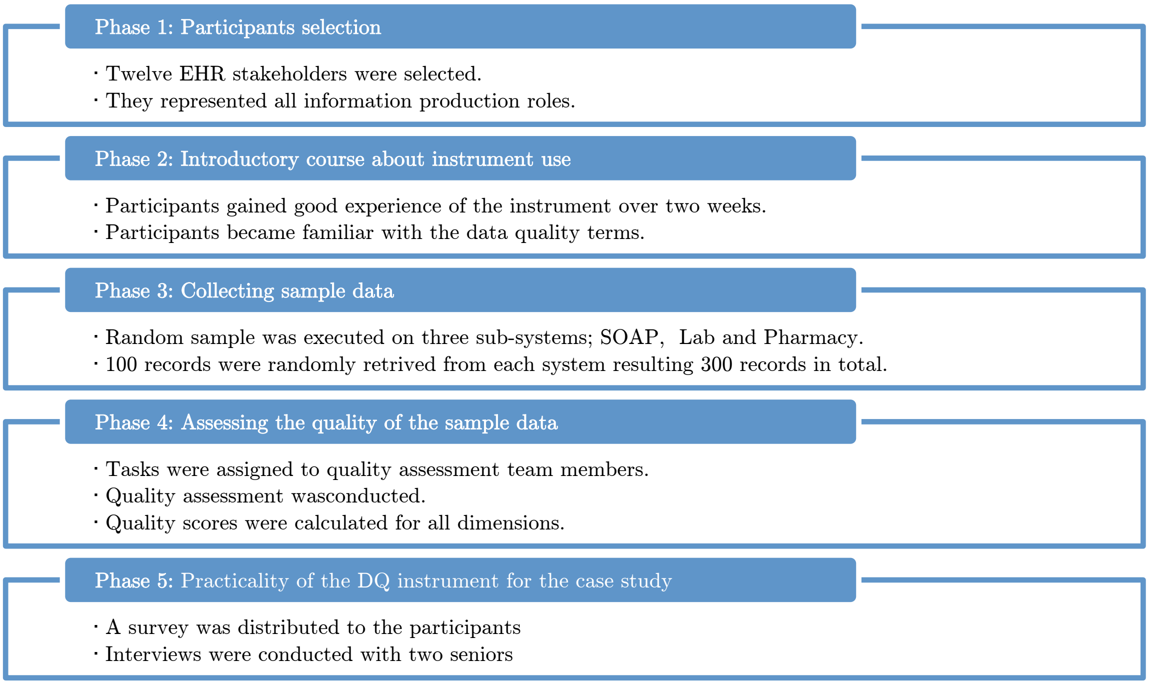
Coding is the process of segmenting and organizing transcripts before bringing meaning to the qualitative data (Rossman & Rallis, 2003). The defined themes (see Figure 4.2) were used to code the data of the focus group. Collating codes into related themes was done in order to gather all data relevant to each theme. After that, the relationships between the themes and coded extracts were checked for accuracy. Subsequently, extracted data in each code were read and analysed in order to establish and confirm the security aspects of data quality and their measures, and relate back the analysis to the research questions. Finally, the findings were checked and a report of the analysis was produced.

4.3.3.6 Ethical Approval

The quantitative and qualitative methodologies conducted in this study were approved by the Ethical Committee of the School of Electronic and Computer Science at the University of Southampton. Ethical approval was granted under reference number **ERGO/FPSE/9382**.

4.3.4 Case Study Design and Procedure

The Section describes the stages and steps involved in conducting the case study. As illustrated in Figure 4.3, the case study consists of five phases including participant selection, an introductory course about DQ instrument use, sample data collection, an assessment of the quality of the sample data and an evaluation of the instrument.



Phase 1: Participants selection

- Twelve EHR stakeholders were selected.
- They represented all information production roles.

Phase 2: Introductory course about instrument use

- Participants gained good experience of the instrument over two weeks.
- Participants became familiar with the data quality terms.

Phase 3: Collecting sample data

- Random sample was executed on three sub-systems; SOAP, Lab and Pharmacy.
- 100 records were randomly retrived from each system resulting 300 records in total.

Phase 4: Assessing the quality of the sample data

- Tasks were assigned to quality assessment team members.
- Quality assessment wasconducted.
- Quality scores were calculated for all dimensions.

Phase 5: Practicality of the DQ instrument for the case study

- A survey was distributed to the participants
- Interviews were conducted with two seniors

Figure 4.3 Case study phases

Phase 1: Five experienced medical practitioners and three health informatics (medical records) staff formed as a team to determine the quality levels of the sample data. They utilized the DQ instrument measures to measure the sample data quality through the objectivity dimensions (accuracy, completeness, consistency and timeliness). They also assessed the quality of the system utility (usability, relevance, interpretability and provenance) using the sample and live system data. Later on, four experienced IT specialists joined this quality assessment team to evaluate the quality of their system security through the integrity-related dimensions[11] (integrity, confidentiality and privacy).

Phase 2: Prior to the assessment process, the quality assessment team was given a two-week introductory course on how to utilize the DQ instrument. This was to train them on the use of the proposed instrument and to assure their capability of achieving the task properly. This course was essential to encounter any resistance that the researcher might confront due to the wrong perception about the main goal of the work.

Phase 3: The sample of 300 patients' records was taken from three sub-systems of the integrated EHR system, namely SOAP (subjective, objective, assessment, and plan), Laboratory and Pharmacy systems.

[11] This category's name was changed later to security-related dimensions to differentiate it from the emerging dimension (integrity) in Chapter 7.

Phase 4: The quality assessment team was exposed to the sample data retrieved from their system to assess. They needed to go through the measures of each dimension in order to examine the quality of the data against each dimension. The DQ instrument has both objective and subjective assessment.

Phase 5: The practicality of the framework and its DQ instrument are to be investigated through a questionnaire and two interviews. The questionnaire was distributed to the quality assessment team to evaluate the DQ instrument through four elements; 'the ease of the use', 'the perceived usefulness', 'the user satisfaction' and 'perception of congruence between expectation of the use and its actual performance'. The interviews were conducted with two senior managers.

4.3.4.1 Context of Study and Participants

The case study was conducted in a large hospital serving a population of almost 400,000 patients. The patients' records are electronically stored in their integrated EHR system. It consists of four subsystems. The hospital regulations allow the researcher to conduct the case study on only three subsystems; SOAP (subjective, objective, assessment, and plan), Laboratory and Pharmacy systems. A sample of 100 records from each system was randomly selected, resulting 300 records in total for the case study.

As shown in Table 4-8, the quality assessment team of 12 people was selected to represent all roles in information production; that is, information collectors, information consumers and IT professionals. They are four IT specialists, three people from the health informatics department, and five consultants and medical staff. Subsequently, they undertook an introductory course on how to use the data quality instrument to assess their system data. Next, a sample of 300 patient records was retrieved from their EHR system for the assessment purpose.

Table 4-8: Selected quality assessment team members

Participant	Position	Experience (years)	Role
P1	Paediatric consultant	+15	Information consumer
P2	Emergency consultant	+10	Information consumer
P3	Neurology consultant	+5	Information consumer
P4	Nursing staff	+10	Information collector
P5	Emergency consultant	+10	Information consumer
P6	Medical records staff	+5	Information collector
P7	Medical records staff	+10	Information collector
P8	Medical records staff	+5	Information collector
P9	Head of IT department	+15	IT professional
P10	System analyst	+10	IT professional
P11	Senior programmer	+5	IT professional
P12	Security expert	+10	IT professional

4.3.4.2 DQ instrument practicality survey

Once the framework of data quality and its items were developed, the practicality of the DQ instrument derived from the framework needed to be examined as a feasibility test for the proposed framework and the defined measurement approach. A questionnaire survey was needed to capture the perceptions of the quality assessment team about the DQ instrument and whether they would continue using this DQ instrument for data quality assurance purposes.

By synthesising the constructs addressed in the TAM and ECM models (see Section 2.4), Table 4-9 summarizes theses constructs and their operational definition (the questionnaire is provided in Appendix D).

Table 4-9. Operationalisation of constructs

Construct	Operational Definition
Perceived usefulness	The degree to which a DQ team member believes in the ability of DQ instrument to measure the quality of their data.
Perceived ease of use	The degree to which a DQ team member believes the use of DQ instrument does not require more effort.
Satisfaction	A DQ member's affect with prior DQ instrument use.
Confirmation	A DQ member's perception of congruence between expectation of the DQ instrument use and its actual performance.

4.3.4.3 Ethical Approval

The Data Quality Assessment case study was approved by the Ethical Committee of Electronic and Computer Science at the University of Southampton, in that the study met the required ethical standards. The approval was granted under reference number **ERGO/FPSE/16276**.

4.4 Chapter Summary

In this chapter, the research methodologies used in this research for the confirmation of proposed 11-dimensional data quality framework and their quality measures are explained. A questionnaire instrument was used to gather data from EHR stakeholders across Saudi Arabia. Semi-structured interviews were conducted with experts and data consumers. Moreover, the revision of the integrity-related dimensions and their measures used the focus group method to review the proposals with security experts in the field as recommended in Sections 6.2.9, 6.2.10 and 6.2.11. After the confirmation of the proposed framework and its associated measures, a case study is used to demonstrate our data quality framework and its data quality (DQ) instrument, to understand the usefulness of the DQ instrument and to examine its practicality and applicability. The findings and the discussions are provided in the next three chapters, while the results of the case study are presented in Chapter 1.

Chapter 5 Confirmatory study Results and Findings

This chapter focuses on electronic health record system (EHR) stakeholders' perspectives with regard to data quality. First, the dimensionality of data quality is examined through a questionnaire. This questionnaire was distributed to and collected from National Health Service (NHS) staff in Saudi Arabia in order to confirm the determinants that represent data quality in EHR in Saudi Arabia.

Semi-structured interviews with experts and data consumers were conducted to ascertain the findings of the questionnaire with regard to the data quality framework. The interviews were also to confirm and generate more items representing the 11 determinants. The experts group consisted of senior database administrators (DBA), physician team leaders and an application analyst. The data consumers group included paediatric consultants, a radiology consultant, an emergency consultant, a medical director and a senior nurse.

The results of the questionnaire were analysed using SPSS software, while NVivo10 helped the researcher to analyse the interviews.

5.1 Results of the Questionnaire

Stakeholders play an important role in establishing the definition of quality in their systems. This became clear when looking at the numerous definitions of quality provided in the literature, maintaining that quality is defined by the customer (Beamon & Ware, 1998; Stracke & Hildebrandt, 2007). Therefore, data quality is context-dependent and the users' perspective is necessary for quality measurement. So the identified data quality determinants were arranged in questionnaire format to capture a user perspective of the relative importance of data quality determinants in EHR systems. This questionnaire survey was carried out over three months from August 2013 with the involvement of 66 clinical and non-clinical staff working for the NHS in Saudi Arabia. It jointly with the semi interviews detailed in Section 5.2 aimed to answer the research question:

RQ1: What data quality determinants are important for EHR stakeholders perceived data quality?

As shown in Figure 5.1, clinical staff members involved in this survey constituted 91% of the respondents: 65% of the participants were clinicians, including GPs and hospital doctors; 17% of respondents were nurses; and other clinical staff formed 9%. Non-clinical staff included a hospital director, and IT-related and finance staff.

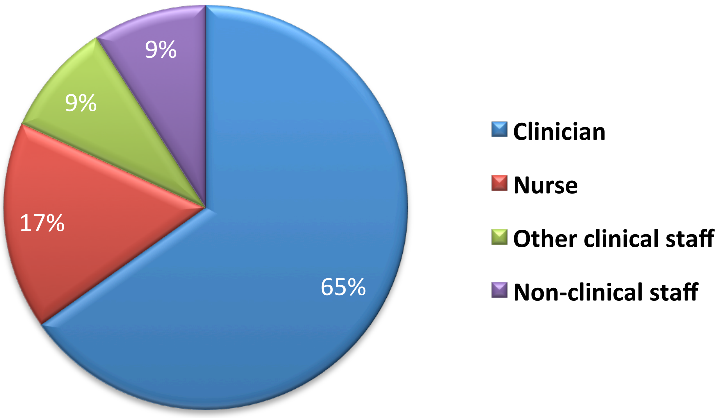


Figure 5.1: Participant careers within the sample

As displayed in Figure 5.2, there was agreement on the importance of accuracy, completeness, consistency, timeliness and usability. No participants registered any objection to the importance of these data quality determinants, as mentioned earlier. On the other hand, secure access was accorded less importance as 5% of the respondents ranked it as 'not important' to data quality assessment, and 9% were 'neutral'. Privacy was also seen less important compared with other determinants, as 4% of the participants believed it was 'not important', while 6% stood in between.

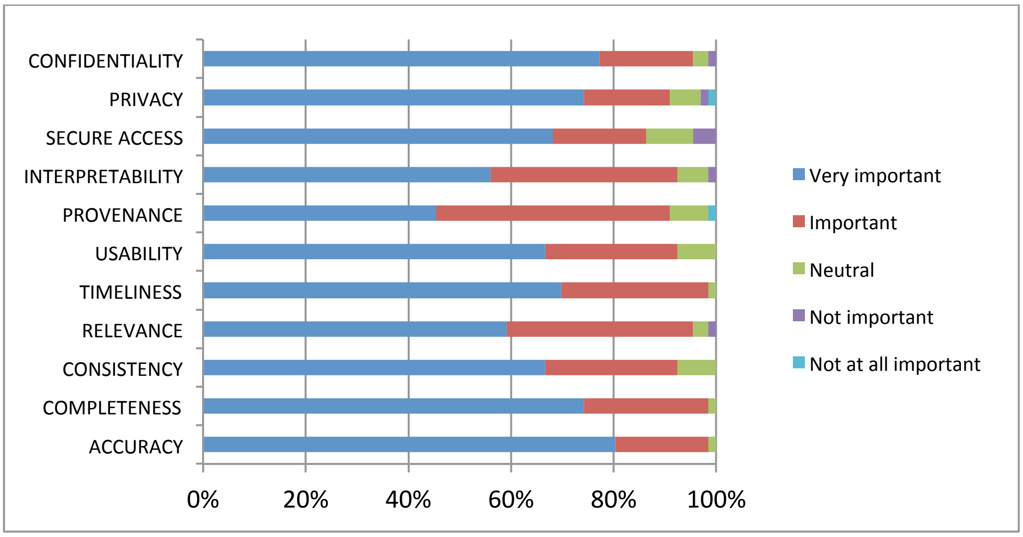


Figure 5.2: The participants' perspectives on the importance of each dimension

As can be seen from Figure 5.3, mean values for each dimension for clinicians, other clinical and non-clinical staff were computed for comparison for these groups of users. It was clear that the two groups of clinicians and non-clinical staff considered these dimensions more important than the other clinical groups, excepting completeness and secure access for non-clinical staff.

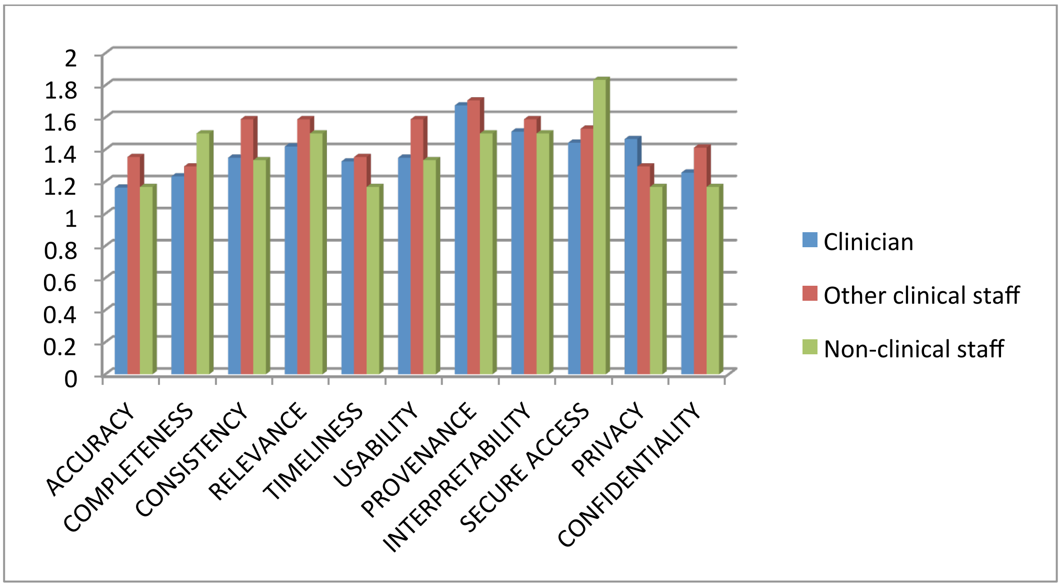


Figure 5.3: Mean values for clinician, other clinical and non-clinical staff

The questionnaire data were analysed using SPSS software to identify the important attributes in EHR systems and to build up a quality framework. The hypothesis was tested using One-sample T test. In the analysis, the sample was compared against the test value (2.5) from the five-point Likert scale ranging from 'very important' (=1) to 'not at all important' (=5).

As shown in Table 5-1, the analysis results show that everyone agreed on the importance of the proposed dimensions as the mean value of each dimension was less than the test values (2.5). The fact that all answers were significant, as *p* values for all dimensions were less than 0.05, confirms that all proposed dimensions were important to the data quality framework in EHR systems.

Table 5-1: One-sample statistics for all dimensions

Dimensions	N	Mean	Std. Deviation	Std. Error Mean	Sig. (2-tailed)
Accuracy	66	1.2121	.44773	.05511	< 0.001
Completeness	66	1.2727	.48184	.05931	< 0.001
Consistency	66	1.4091	.63190	.07778	< 0.001
Relevance	66	1.4697	.63778	.07850	< 0.001
Timeliness	66	1.3182	.50105	.06167	< 0.001
Usability	66	1.4091	.63190	.07778	< 0.001
Provenance	66	1.6667	.75107	.09245	< 0.001
Interpretability	66	1.5303	.68432	.08423	< 0.001
Secure access	66	1.5000	.84580	.10411	< 0.001
Privacy	66	1.3939	.80151	.09866	< 0.001
Confidentiality	66	1.2879	.60167	.07406	< 0.001

In response to the question '*Is there any other characteristic not mentioned earlier that needs to be considered?*', there were four proposals for the data quality framework. One of the IT staff working for the National Health Ministry proposed *integrity* and *availability* as necessary dimensions for the framework. Likewise, a hospital doctor believed that *user friendly* is a requirement for the data quality framework, and a senior nurse suggested *system security* as a further dimension.

It was noticed that two of the respondents considered that the dimension of 'provenance' was unclear and needed clarification. Apart from provenance-related comments, all dimensions were clear and understandable. (Refer to APPENDIX B to see the results in details)

5.2 Interviews Findings and Results

As described in the research methodology chapter, this Section focuses on the expert and data consumer perspective with regard to data quality dimensions and the quality measures associated with each dimension. Semi-structured interviews were carried out with several experts and EHR stakeholders in the health sector. The main aim of the interviews was to examine the dimensionality of data quality as well as the reliability of multidimensional scale identified. These interviews aimed to partially answer the research questions:

RQ1: What data quality determinants are important for EHR stakeholders perceived data quality?

RQ2: Based on the proposed framework, what is the appropriate instrument with which to measure quality of data in EHRs?

RQ3: what are the severity factors that make data quality problems more severe?

Interviews were conducted with two groups: experts and data consumers. The expert group consisted of two senior database administrators, two physician team leaders and an application analyst. The other group included a medical director, two paediatric consultants, an emergency consultant, a radiology consultant and a nurse with 13 years' experience.

The results were analysed using NVivo10 software. Codes were developed from the interviewees' answers and dimensions, to capture the effectiveness of a dimension to measure the quality of data. The codes include the relevance of an associated item to this particular dimension, and the factors that make quality problems even more severe.

5.2.1 The Expert perspective

This Section will briefly describe the findings and results of interviews conducted with the expert staff members responsible for EHR maintenance and assuring the high quality of data in their systems. Interviews were conducted with five experts in order to examine the potential of the proposed measures and its quality items, and the possibility of enhancing the measures by tackling more aspects of data quality problems not covered in the proposed measures. After analysing the semi-structured interviews with experts, the output was themed into:

- 11-dimensional data quality framework (RQ1)
- Data quality measures (RQ2)
- Interpretation of the assessment results (RQ3)

5.2.1.1 11-dimensional data quality framework

The data quality scale was discussed in detail with experts with the aim of reaching a valid set of dimensions that properly assesses the quality of data populating the EHR systems. Our 11-dimensional framework was examined from different angles. Experts were asked what data quality determinants (dimensions) were important for data quality assessment in EHR systems, and whether the proposed 11-dimensional framework was adequate for reliable assessment of the quality of data. The phrases relating to this matter were grouped into three main themes; dimension importance, definition, and the adequacy of our 11-dimensional scale.

With regard to the importance of dimensions, all experts involved in this interview agreed that each dimension in our scale is essential to the overall quality of data. The phrases showed highly positive attitudes towards these dimensions and their role in the assessment stage. It was noticeable that the most common terms used in their expressions were 'important', 'essential' and 'no question about it'.

Regarding the definitions pf dimensions, there were concerns with three. First, a team leader felt that the definition of consistency needed some attention to cover more aspects. The other dimension was privacy definition, and a senior DBA was concerned that systems should protect individuals, not the other way around. The third was the most controversial point, and concerned the completeness dimension; four of the experts rejected the second part of the definition. The following quotes are provided in support of this concern:

Team Leader1: *For me, the definition is OK, but needs some attention to cover all aspect of consistency.*

Senior DBA1: *The wording of the definition of privacy needs to be looked at as the system should seclude individuals, not the other way around.*

Senior DBA2: *First part of definition is clear and measurable. But the second, I am not sure whether is relevant to completeness in the context of data quality.*

Team Leader2: *First part of definition is sound and clear, and reflects my understanding of completeness. However, the second part, my point of view, has nothing to do with completeness and depends on the interpretation of the user.*

Team Leader2: *First part does reflect, but I am not sure about the second part.*

Analyst: *The first part of the definition is sound and clear. Yet, the second one is not clear and its exclusion is recommended as the first bit covers the aspects of completeness.*

Apart from the feedback mentioned earlier, all experts were satisfied with the definitions assigned to the dimensions. They confirmed that all definitions reflected their understanding of dimensions and their implications.

Finally, the experts were asked for our 11-dimensional scale to be tested for adequacy and comprehension. In response, nothing was proposed; they believed that the identified dimensions were good indictors for the quality of data in EHR systems. Quotes below support this:

Senior DBA1: *I guess this framework covers valuable and important aspects needed to leverage the quality of data in health information systems. However, the selections of database management systems (DBMS) should be given a great attention as some DBMS support data quality activity.*

Team Leader2: *I guess this framework is comprehensive and covers all aspects needed to improve the quality of data in health information systems.*

5.2.1.2 Data quality items

In this Section, the proposed measures were discussed with experts in order to confirm their relevance and to explore more items that address quality problems as yet not covered by the measures. The results and findings of the interviews were categorised into confirmed; irrelevant; overlapping; and additional items for each dimension.

- ***Accuracy***

As discussed in the literature review, there are seven items that fall into the category of accuracy assessment, according to the implication of the definition of accuracy. These, shown in Table 5-2, have been through preliminary refactoring and classification, and ended up under the accuracy measures. This would allow objective assessment of the quality of accuracy.

Table 5-2: The proposed measures for accuracy

Item code	Data quality item	Example
ACC1	Illegal values due to invalid domain range	DoB= 30/02/1980
ACC2	Misspellings	Country= Germiny
ACC3	Misfielded values	City= UK
ACC4	Embedded values	Name= ‘J. Smith 12-01-2011 London’
ACC5	Word transposition	Name1= ‘J. Smith’ name2= ‘Muller K’
ACC6	Wrong reference	Deptno= 12, found but incorrect
ACC7	Erroneous entry	Age= 26, wrong real age= 36

After analysing the semi-structured interviews with the experts, all experts confirmed that the proposed quality items are sound measures and relevant to accuracy. They believe that these items do not overlap others within the measures.

In addition, two of the experts noted some problems that could compromise the quality of accuracy. Their two suggestions are quoted below:

Senior DBA1: *In data integration, orphan information affect the accuracy of information, this is due to the absence of primary-key and foreign-key relation.*

Senior DBA2: *Data validation needs to be considered.*

- ***Consistency***

To arrive at valid measures for consistency assessment of EHR data, interviews were conducted with experts who have been dealing with health databases for more than 10 years. Table 5-3 shows the quality items drawn from the literature review. These were discussed intensively with the experts to refine them and to address problems not covered by the measures.

Table 5-3: The proposed measures for consistency

Item code	Data quality item	Example
CON1	Violated attribute dependencies	Age value not compatible with DoB value
CON2	Uniqueness violation	Uniqueness of student ID violated
CON3	Naming conflicts in multi-source	
CON4	Structural conflicts in multi-source	
CON5	Wrong categorical data	Not user-specified terms
CON6	Referential integrity violation	DeptNo=11, not found
CON7	Violated attribute dependencies	Postcode=SO171BJ city=London
CON8	Duplicated records in single/multi data source(s)	A student has 2 records
CON9	Contradicting records in single/multi source(s)	Same entity described differently in 2 records
CON10	Inconsistent spatial data	
CON11	Different measure units in single/multi source(s)	
CON12	Syntax inconsistency	2 dates with different format

The consistency quality items underwent a process of examination and refactoring to achieve a list of relevant and reliable measures for consistency measurement.

Table 5-4 highlights the outcome of the interviews analysis.

Table 5-4: Experts' findings for consistency measures

Item code	Senior DBA1			Senior DBA2			Team Leader1			Team leader2			Analyst		
	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O
CON1	✓			✓			✓			✓			✓		
CON2	✓			✓			✓			✓			✓		
CON3	✓			✓			✓			✓			✓		
CON4	✓			✓			✓			✓			✓		
CON5		✓	✓		✓		✓			✓			✓		
CON6		✓			✓		✓			✓				✓	
CON7	✓					✓	✓			✓			✓		
CON8	✓			✓			✓			✓			✓		
CON9		✓			✓		✓			✓			✓		
CON10		✓			✓		✓			✓			✓		
CON11	✓			✓			✓			✓			✓		
CON12	✓			✓			✓			✓			✓		

C: Confirmed item	I: Irrelevant item	O: Overlapping item

The two physician team leaders confirmed that all quality items are relevant and sound measures for consistency. However, the two senior DBAs and an application analyst claimed that the item '*referential integrity violation*' is not a consistency-related measure, but is accuracy-related. Moreover, the two DBAs considered the items CON5, CON9 and CON10 were sound measures of accuracy. Senior DBA1 claimed that the items of '*wrong categorical data*' and '*referential integrity violation*' overlap, as they address the same issue.

It is worth noting that the experts emphasised that the measures of consistency are comprehensive and cover the required aspects to measure this dimension. The following examples support this argument:

Analyst: *I guess these items would properly assess the consistency.*

Team Leader2: *I think these items cover all aspects of consistency that we have in my mind.*

- ***Completeness***

The quality items associated with completeness dimension were examined by interviews with experts in the health field, and Table 5-5 displays those proposed.

Table 5-5: Proposed measures for completeness

Item code	Data quality item	Example
COM1	Missing data where Null-not-allowed constraint enforced	
COM2	Missing data where Null-not-allowed constraint not enforced	Data was unknown during initial stage
COM3	Missing record	A case not entered
COM4	Ambiguous data due to incomplete context	
COM5	Semi-empty tuple	

The phrases relating to quality items associated with completeness were positive responses made when these experts were giving their opinion regarding the proposed measures. All confirmed its relevance to the completeness dimension and its comprehensiveness in covering all required aspects of this dimension. The following examples support the findings mentioned earlier:

Analyst: *I can't think of anything to contribute to these items. I guess it is comprehensive.*

Team Leader2: *I have nothing to add, and I think it is covered all aspects of completeness.*

- ***Timeliness***

To address this dimension, interviews were carried out with experts to explore what aspects needed to be considered in order to measure the quality of timeliness. Initially, two proposed items drawn from data cleaning studies: 'Outdated temporal value' and 'Outdated reference'.

Considering all phrases relating to the proposed items for timeliness assessment, the responses were almost entirely positive, aside from the second item. The two DBA experts consider that item ineffective as the DBMS avoids such problem by update cascading. No overlap was detected within these measures and the experts did not add any items, believing it was sufficient:

Analyst: *I think these items cover the aspects needed to judge the quality of timeliness.*

- ***Interpretability***

There was a mixture of opinions on the proposed interpretability-related measures.

Table 5-6 includes the quality items that were an important subject of discussion with the experts. There was no clear agreement on the proposed measures, and many issues were raised during discussion.

Table 5-6: Proposed measures for interpretability

Item code	Data quality item	Example
INT1	Different representations due to use of abbreviation and cryptic values	
INT2	Different representations due to use of Alias/nickname	
INT3	Different representations due to use of encoding format	ASCII
INT4	Different representations due to use of special characters	Space, no space, dash

As shown in Table 5-7, the experts agreed that the first item is a measure for interpretability, and almost all agreed on the last item. However, their quotes did not address the different representations due to use of abbreviations and cryptic values but the use of them. Most of them also agreed that the item '*different representations due to use of encoding format*' is irrelevant to interpretability as it is associated with the consistency measures. Furthermore, four out of five experts believed that using aliases or nicknames is not applicable in health systems in Saudi Arabia, and therefore did not consider it a relevant measure.

Table 5-7: Experts' findings for interpretability measures

Item code	Senior DBA1			Senior DBA2			Team Leader1			Team Leader2			Analyst		
	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O
INT1	✓			✓			✓			✓			✓		
INT2		✓			✓			✓			✓		✓		
INT3		✓			✓		✓				✓			✓	
INT4	✓				✓		✓			✓			✓		

C: Confirmed item | I: Irrelevant item | O: Overlapping item

Apart from these proposed items, respondents added no items for interpretability measurement since they believed it to be a subjective dimension.

- ***Usability***

The usability measures, as shown in Table 5-8, were discussed with a group of experts to reach a valid and reliable consensus. Opinions were a mix of positive and negative answers on the proposed quality items.

Table 5-8: The proposed measures for usability

Item code	Data quality item
USA1	Is this information easily accessible
USA2	Is this information easily retrievable
USA3	Is this information not usable
USA4	Is this information promptly accessible when needed
USA5	Is this information easily updatable

USA6	Is this information easily understood
USA7	Is this information usable
USA8	Is the availability of information for the patient treat adequate
USA9	Is this information easily manipulated

They almost all agreed on five quality measures, illustrated in Table 5-9, and believed in their reliability for assessing the quality of data usability. These are USA1, USA2, USA4, USA8 and USA9, and most objected to USA3 and USA7 as generic and not measurable.

Table 5-9: Experts' findings for usability measures

Item code	Senior DBA1			Senior DBA2			Team Leader1			Team Leader2			Analyst		
	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O
USA1	✓			✓			✓			✓			✓		
USA2	✓			✓			✓			✓	✓		✓		
USA3		✓			✓			✓		✓	✓		✓		
USA4	✓			✓			✓			✓			✓		
USA5			✓			✓	✓			✓			✓		
USA6	✓			✓				✓			✓		✓		
USA7		✓			✓			✓		✓			✓		
USA8	✓			✓			✓			✓			✓		
USA9	✓			✓				✓		✓			✓		

C: Confirmed item | I: Irrelevant item | O: Overlapping item

Furthermore, the experts were not in agreement over two aspects of data usability; data being understandable (USA6) and updatability (USA5). The two team leaders argued that the former should be moved into the interpretability dimension, while the two DBAs assumed that the latter overlapped with the manipulation feature (USA9).

- ***Relevance***

Considering all the phrases relating to the proposed quality items displayed in

Table 5-10, responses were positive and reflect their functionality and practicality in measuring the quality of relevance.

Table 5-10: The proposed measures for relevance

Item code	Data quality item
REL1	Is this information relevant to the task at hand
REL2	Is this information useful to the task at hand
REL3	Is this information applicable to the task at hand
REL4	Is this information appropriate for the task at hand
REL5	Is this information irrelevant to the task at hand

However, three of the experts proposed excluding the two items that use the word 'relevant' (REL1 and 5), as this is generic and is derived from the dimension name. The following example supports this point:

Senior DBA2: *My point of view, we should not use the word 'relevant' as we try to find items to measure relevance.*

With regard to additional aspects, respondents think these measurable items are sufficient to cover all the aspects needed to assess the quality of relevance. The following example reflects their opinion:

Analyst: *I reckon that these items are sufficient to assess the relevance of the information, and no more items are needed.*

- ***Provenance***

Phrases related to the functionality of the proposed set of items to measure the quality of provenance were mainly positive. Respondents agreed that the four aspects needed to assess the provenance of data are: the origin of data, the authorship, the creation date and the update history.

- ***Secure access***

Although the experts approved most of the proposed items for the secure access dimension, they concluded their discussion by noting that there is much similarity between these items. Some suggested that these measures need to be reviewed by security experts in order to reach a reliable set of measurable items for the dimension of secure access. The following quote represents this view:

Senior DBA1: *We need to give special care for security as it is the safeguard of our data. Security experts should be involved into this stage.*

However, some thought that these measures are sufficient to assess this dimension since it involves the measurement of authentication and authorisation that comprise the two important aspects of access control.

Table 5-11: The proposed measures for secure access

Item code	Data quality item
ACC1	Is the access to this information sufficiently restricted
ACC2	Does the access to this information require authentication process
ACC3	Is the owner/creator of this information authorised
ACC4	Is unauthorised access to this information sufficiently prevented

Phrases relating to proposed measurable items showed that there is general agreement on the functionality of the measures, apart from the last item that two experts considered to overlap with item 1; see Table 5-11.

- ***Confidentiality***

According to interviewees' phrases on confidentiality, only two quality items were a subject of debate. Therefore, apart from CONF6, which was believed by the application analyst to be a redundant copy of item CONF4, the remainder was agreed; see Table 5-12:

Table 5-12: The proposed measures for confidentiality

Item code	Data quality item
CONF1	Is the confidentiality of this information achieved
CONF2	Is this information only accessible by authorised people
CONF3	Is access to this information granted only to persons who should see it
CONF4	Is this information vulnerable
CONF5	Is the sensitivity of this information clearly declared
CONF6	Could this piece of information be disseminated without permission
CONF7	In case of sharing, is there a clear consent for this information to be shared

As shown in

Table 5-13, item CONF3 was clearly believed to be a duplicate of item CONF2. This redundancy was noticed by all the experts. Moreover, the senior DBA1 argued that item CONF4 is supposed to be within the measures of secure access, not the confidentiality dimension.

Table 5-13: Experts' findings for confidentiality measures

Item code	Senior DBA1			Senior DBA2			Team Leader1			Team Leader2			Analyst		
	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O
CONF1	✓			✓			✓			✓			✓		
CONF2	✓			✓			✓			✓			✓		
CONF3			✓			✓			✓			✓			✓
CONF4		✓		✓			✓			✓			✓		
CONF5	✓			✓			✓			✓			✓		
CONF6	✓			✓			✓			✓					✓
CONF7	✓			✓			✓			✓			✓		

C: Confirmed item | I: Irrelevant item | O: Overlapping item

- ***Privacy***

As shown in Table 5-15, phrases about privacy-related items form a pattern. These responses confirmed the functionality of the proposed items apart from items PRI3 and PRI4, which are believed to overlap with item PRI2.

Table 5-14: The proposed measures for privacy

Item code	Data quality item
PRI1	Is personally identifiable private information being appropriately safeguarded
PRI2	Is this personally identifiable information compliant with privacy policy
PRI3	Is the policy regulation obeyed in this piece of information
PRI4	Is privacy policy violated in this piece of information
PRI5	Is the access to this personally identifiable information granted only to persons who should see it
PRI6	In case of an individual being identified, is there a clear consent from this individual to be identified

However, the senior team leader1 came up with a slightly different response. He went against the relevance of item PRI5 to the privacy measurement, since he believed it is secure access-related. Moreover, he felt that there was no overlap between items PRI2 and PRI3.

Table 5-15: Experts' findings for privacy measures

Item code	Senior DBA1			Senior DBA2			Team Leader1			Team Leader2			Analyst		
	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O
PRI1	✓			✓			✓			✓			✓		
PRI2	✓			✓			✓			✓			✓		
PRI3			✓			✓	✓					✓			✓
PRI4			✓			✓			✓			✓			✓
PRI5	✓			✓				✓		✓			✓		
PRI6	✓			✓			✓			✓			✓		

C: Confirmed item | I: Irrelevant item | O: Overlapping item

5.2.1.3 Interpretation of the assessment results

The interpretation process follows the assessment stage. It is a quality-reasoning that uses metadata produced at the assessment stage with additional inputs such as weightings and user preferences. The phrases concerning the severity factors that make the quality even worse show responses fairly rich with some important issues.

Table 5-16 summarises the findings with regard to severity factors:

Table 5-16: Severity factors proposed by experts

Severity factor	Proposer
Error distribution	Senior DBA1
Data sensitivity	Senior DBA1, Team Leader1
Data quality items	Senior DBA2
Type of data collection tools	Team Leader1

During interviews, once a factor was proposed it was put before the next interviewee to examine his/her opinion about it. All experts confirmed the impact of these factors on the quality problem, meaning that the factors make it even worse. The following quotes support the severity factors:

Senior DBA1: *Error distribution is one of the issues that make problem fixing difficult and even harder. Sensitivity of data is another factor that worsens the impact of errors on this data.*

Senior DBA2: *Error distribution could make the quality problems complex as it requires more effort and time. We should consider the type of error. The impact of erroneous entry is not like misspelling errors.*

Team Leader1: *My point of view, the factors are as follows:*

- *Error distribution*
- *Type of data that holds problem, problems found in clinical data are not like those found in administrative data.*
- *Type of data collection that generates errors. For example, if an error is generated by a machine, we need to check all data generated by this machine, which consume data and effort.*

Furthermore, the IT professionals' responses to the questions, 'Does the importance of this dimension vary along with the task at hand?' and 'Does the importance of this dimension vary along with the class of user at hand?' showed that a dimension is not affected by having different tasks and classes of user. However, all data consumers perceived that some dimensions are more important than others. Moreover, the measures within one dimension are not equally important. This finding was gained through the interviewees' responses of the question, 'from your point of view, do these measuring items have the same weight in assessing the information in terms of "dimension"?'

5.2.2 Data Consumers' Perspective

This Section will briefly cover the findings and results of interviews conducted with data consumers in the health sector, namely a medical director, four consultants and an experienced nurse. The contribution of EHR stakeholders to this research is crucial since they consume data on daily basis and could provide us with problems regarding the quality of EHR data. After analysing the semi-structured interviews, the output was themed as follows:

- 11-dimensional data quality framework (RQ1)
- Data quality measures (RQ2)
- Interpretation of the assessment results (RQ3)

5.2.2.1 11-dimensional data quality framework

This Section highlights the data consumers' feedback with regard to the identified determinants that make up the domain of the data quality framework. The interviews covered the definition of each determinant, its importance and whether the proposed 11-dimensional framework is adequate to provide reliable assessment of the quality of data. The phrases relating to this matter were grouped into three main themes: dimension importance, definition and the adequacy of our 11-dimensional scale.

With regard to the importance of dimensions, all data consumers involved in the interviews agreed that each dimension of our scale is essential to the overall quality of data. The phrases showed highly positive attitudes towards these determinants and their role in assessment stage. It was noticeable that the most common terms used in their expressions were '*important*', '*essential*', '*pillar of data quality*' and '*no question about it*'.

Regarding the definitions of determinants, there was agreement on seven determinants. However, some professionals had concerns over the definitions of four. First, the medical director argued that the definition of accuracy would not be clear to end users. He suggested that the term 'actual values' should instead be 'documented values'. Secondly, the emergency consultant proposed omitting the second part of the completeness definition. Most controversial was the definition associated with secure access dimension, as three data consumers were opposed to the use of the word 'personal'. They said that all data, not only personal, should be secure. The last was the privacy-related definition. The medical director and paediatric consultant said the definition needed attention. The following quotes support these various claims:

Med Director: *For normal people, it is a bit hard to absorb this definition. I understand it well, but I guess the phrase 'actual value' needs to be replaced with something like 'documented value'.*

Consult4: *First part of definition is clear and sufficiently reflective of my understanding of completeness. However, the second one is extra to the definition of completeness, and should be omitted.*

Consult1: *As all patients' data should require secure access, the definition does only include personal data. EHR data include all patients-related data whether it is personal or clinical. So I recommend the word 'patient' as a replacement of 'personal'.*

Consult2: *This definition implies that this dimension is only for personal data. 'Personal data' should be replaced with 'patient data' which then include all patient-related data.*

Med Director: *This definition needs a little attention. I think the word 'personal' or 'patient' should be removed leaving the word 'data' without adjective, as EHR data include personal, clinical and patient data.*

Consult3: *I guess this definition needs to be clearer.*

Med Director: *This definition needs attention.*

Apart from these points, all data consumers were satisfied with the definitions given to the determinants. They confirmed that all reflected their understanding of determinants and their implications.

Finally, some exploratory questions were asked to judge the sufficiency of our 11-dimensional scale, to add more determinants if some aspects of data quality were missing. Nothing was proposed in their responses, and experts believed that the identified determinants constituted a firm framework for the quality of data in EHR systems. Sample quotes are provided:

Consult4: *I guess these dimensions constitute a firm framework for data quality assurance and all what I can think of is already covered.*

Nurse: *I guess the dimensions mentioned in the framework are all important quality aspects I can think of.*

5.2.2.2 Data quality items

In this Section, the proposed items that represent each determinant are discussed with data consumers in order to confirm their relevance and explore more items that address quality problems not covered within the measures. The results and findings of the interviews were categorised into confirmed, irrelevant, overlapping or additional.

- ***Accuracy***

The proposed measuring items of accuracy were discussed with experienced data consumers. All responses were positive, indicating that all items were agreed upon as sound measures relevant to accuracy. Interviewees also denied that there was any overlap of items.

In response to the question '*Are there any other quality items that need to be considered?,*' most interviewees confirmed that this set of measures is sufficient for accuracy measurement. Only one physician added: '*As a physician, we need to measure accuracy in lab values and x-ray readings, or any results that directly affect a patient treatment.*'

- ***Consistency***

In spite of the fact that experts have many issues with regard to items that assess the quality of consistency, the data consumers almost all agreed on them being good measures relevant to consistency. Moreover, they added that there is no overlap of the quality items. However, a paediatric consultant claimed that '*contradicting records*' and '*duplicated records*' addressed the same problem, being caused by duplication. Another paediatric consultant did not consider the item '*inconsistent spatial data*' as applicable to the EHR domain. The following quote supports this point:

Consult1: *I am not sure whether it is applicable to our area.*

It is worth noting that a medical director and an emergency consultant shed light on an annoying problem with dual-language names. They expressed concern at the absence of a standard or protocol on translating Arabic names into English. This problem gives rise to different spellings for the same name. Here is the emergency consultant's quote:

Consult4: *We have a noticeable consistency problem with patients' names as one family name could have many spellings in English. Many fatal incidents occurred here were caused by such consistency problems.*

- ***Completeness***

Considering the phrases related to the completeness measures, the responses showed agreement on its functionality and relevance; experts and data consumers had the same response towards the proposed quality items of completeness.

No one offered any additional items as they were satisfied with the measures' comprehensiveness and coverage. The followings are examples of the responses:

Med director: *I have nothing to add to these items. I think these items would provide a good assessment for completeness.*

Consult3: *I think this set of items is good enough to assess the quality of completeness.*

- ***Timeliness***

The phrases related to the items proposed to measure the quality of timeliness were all positive. Two items were believed to be good measures and relevant to timeliness. Furthermore, the respondents said that these measures are good enough to assess the quality of timeliness, so did not propose any additional items to the set of measures:

Consult2: *Vital signs need to be up-to-date and available on time. If out-dated temporal value item covers this, so I guess this dimension is fully covered.*

Consult1: *I guess it covers the corresponding aspects, and I have nothing to add.*

- ***Interpretability***

According to the phrases on interpretability assessment, the proposed measures were not satisfactory for data consumers as many items were seen to be irrelevant and some aspects of interpretability were not covered. Moreover, a consultant believed that any difference in data representation should be considered as a consistency-related item.

Table 5-17: Data consumers' findings for interpretability measures

Item code	Consult1			Consult2			Consult3			Med director			Consult4			Nurse		
	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O
INT1	✓			✓			✓			✓			✓			✓		
INT2	✓			✓				✓			✓		✓				✓	
INT3		✓			✓			✓		✓			✓			✓		
INT4		✓			✓		✓			✓			✓			✓		

C: Confirmed item | I: Irrelevant item | O: Overlapping item

As illustrated in Table 5-17, data consumers emphasised that using ad hoc abbreviations and cryptic values reduces the quality of data being interpreted. With regard to the second item, there was no agreement on whether different representations due to the use of aliases or nicknames is an applicable aspect of interpretability. When considering phrases relating to the use of coding format (INT3), there is an agreement on its functionality as a sound measure for data quality assessment. However, three interviewees claimed that this measure is associated with consistency, not interpretability:

Consult1: *Different encoding formats for the same object affects the consistency.*

Consult2: *Encoding format is a good technique to avoid wrong entries. However, having different encoding formats for the same object leads to inconsistency.*

It is worth noting that two experienced professionals added an aspect that needs to be highlighted. The problem lies when there is a disparity in the interpretation of data. Two quotes support this point:

Consult2: *My concern is how to measure the problem of having different interpretations for the same reading. This dimension needs more attention.*

Consult3: *There is one thing that needs to be considered. That is, different interpretations for the same data. For example, in lab results, there are different acceptable ranges due to different schools of thought. So this needs to be standardised, otherwise we will end up with many interpretations.*

- ***Usability***

Data usability was considered to be one of the most important dimensions needing attention as it facilitates the use of data. Table 5-18 summarises the data consumers' views with regard to its proposed measuring items.

Table 5-18: Data consumers' findings for usability measures

Item code	Consult1			Consult2			Consult3			Med director			Consult4			Nurse		
	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O
USA 1	✓			✓			✓			✓			✓			✓		
USA 2	✓			✓			✓			✓			✓			✓		
USA 3		✓			✓			✓			✓		✓			✓		
USA 4	✓			✓			✓			✓			✓			✓		
USA 5	✓			✓			✓			✓			✓			✓		
USA 6	✓			✓			✓			✓			✓			✓		
USA 7		✓			✓			✓			✓		✓			✓		
USA 8	✓			✓			✓			✓			✓			✓		
USA 9	✓		✓			✓	✓				✓		✓			✓		

C: Confirmed item I: Irrelevant item O: Overlapping item

Most of the data consumers disagreed with the use of the word 'usable' as a measure (USA3 and 7). It was believed to be a generic word derived from the name of its dimension. Some professionals also pointed out that the item USA9 needs more attention as the word 'manipulate' is generic and negative. Some quotes below support this point:

Med Director: *I am not sure about this measure. I guess it covers update item and more.*

Consult2: *The word manipulation is a negative word, it is similar to update.*

Some professionals emphasised the importance of availability as an important constituent that needs to be included in these measures. Here are some quotes for the point:

Med Director: *Availability is an important aspect of data usability. It needs to be considered here. For example, a new drug has been entered into the system to appear in dropdown list at the time of subscribing this drug for a patient.*

Nurse: *It is worth considering that relevant information in different departments should be accessible by authorised persons without the need to send them off.*

- ***Relevance***

When considering all phrases related to the proposed quality items of relevance, data consumers had different opinions, as presented in Table 5-19. Many professionals were against any item that derived its name from the dimension. Apart from this point, there was agreement on the proposed items. The following quotes support the point mentioned earlier:

Consult1: *It (REL1) is overlapped with appropriate and should be excluded as [it] is derived from dimension name.*

Consult2: *It is overlapped with others, especially appropriateness.*

Consult3: *I think REL1 is overlapping with all as it is a broad term.*

Med Director: *Again, this word is broad and generic.*

Table 5-19: Data consumers' findings for relevance measures

Item code	Consult 1			Consult 2			Consult 3			Med director			Consult 4			Nurse		
	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O
REL1			✓			✓		✓			✓		✓			✓		
REL2	✓			✓			✓			✓			✓			✓		
REL3	✓			✓			✓			✓			✓			✓		
REL4	✓			✓			✓			✓			✓			✓		
REL5			✓			✓		✓			✓		✓			✓		

C: Confirmed item | I: Irrelevant item | O: Overlapping item

Furthermore, data consumers claimed that this set of measures would be good enough to provide sound assessment of the quality of data being relevant. Here are some quotes:

Consult3: *These measures are enough to measure relevance.*

Consult2: *This set of measuring items would give us a good assessment for relevance.*

- ***Provenance***

Taking into consideration the phrases relating to the quality items for provenance assessment, all responses were positive, indicating that they are satisfactory and relevant to the provenance dimension. Moreover, all responses indicate that this set of measures covers all important aspects that would properly measure the quality of provenance dimension. So, no additional elements were proposed. The following quotes highlight this finding:

Consult4: *I guess this set of measures covers ownership, origin and update history, which are most important parts of provenance.*

Nurse: *I guess these items could produce a good instrument for provenance.*

- ***Secure access***

After analysing the phrases related to the measures of secure access, most professionals approved the quality items associated with secure access dimension. They believed that the aspects addressed by the proposed measures would produce a valid assessment for secure access. However, two professionals claimed that item ACC4 overlaps with ACC1 since they address the same aspect and have the same implications.

During the interviews, various proposals were suggested to enhance the effectiveness of the proposed measures. A nurse said that some precautions should be taken against illegal practices such as logging off the system if idle for a period. Moreover, a consultant argued that there should be an item to assess data recoverability. Furthermore, the medical director said the set of measures should check whether the remote access policy applies in the event of remote access. He was concerned about this point as he had recently launched a service that allows remote access by professionals.

- ***Confidentiality***

Regarding confidentiality measurement, almost all responses were positive when discussing the associated quality items, aside from item CON3, which was declined by all respondents as it is a rephrased version of item CONF2. Three of the respondents rejected item CON1 for being generic and not measurable. The following quote supports this last point:

Med director: *This item is generic and covers all items mentioned.*

Consult1: *It is a general item, and I am not sure whether it is measurable.*

- ***Privacy***

Considering the proposed quality items associated with privacy, there was a mixture of negative and positive phrases. Table 5-20 summarises the findings drawn from the interviews conducted with the data consumers.

Table 5-20: Data consumers' findings for privacy measures

Item code	Consult1			Consult2			Consult3			Med director			Consult4			Nurse		
	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O
PRI1	✓			✓			✓			✓			✓			✓		
PRI2	✓			✓			✓			✓			✓			✓		
PRI3	✓			✓					✓			✓	✓			✓		
PRI4			✓			✓			✓			✓			✓	✓		
PRI5			✓	✓			✓			✓			✓			✓		
PRI6	✓			✓			✓			✓			✓			✓		

C: Confirmed item | I: Irrelevant item | O: Overlapping item

There was nearly consensus that item PRI4 and PRI3 overlap in addressing the same aspect. In addition, two of the stakeholders claimed that the items mentioned earlier overlap with item PRI2. Moreover, a consultant said that PRI5 is a copy of item PRI1. Other than this, responses confirmed the relevance and the functionality of the measures, and no additions were proposed.

5.2.2.3 Interpretation of the assessment results

Drawing out the severity factors was undertaken in either explicit or implicit ways. The explicit way was a direct question tacking this issue. The implicit way was through analysing professionals' responses, finding phrases that indicate the severity, such as the use of expressive terms such as 'impact' or 'severe'. Table 5-21 summarises the findings with regard to severity factors:

Table 5-21: Severity factors that make data quality problems even worse, as proposed by data consumers

Severity factor	Proposer
Closeness to the correct values	Consult3
Cascading errors	Consult2
Data quality items	Senior DBA2
Data sensitivity	Team Leader1
Type of data collection	Team Leader1

During interviews, once a factor was proposed it was put before the next interviewee to examine his/her opinion about it. Even the severity factors proposed by the experts were put before the data consumers for discussion. The conclusion was that all factors were approved for having impact on quality problems. Some samples of the responses are as follows:

Consult3: *But in demographic data, it depends on the data holding that has the error and the how much the difference between the reference data and the given data. For example, if the actual value of patient's age is 50 and was given 51, I might tolerate such mistake, but if the difference was 20 years old, this is a critical issue.*

The closeness to the correct values helps us decide whether to tolerate this error.

Consult2: *The only thing I can add is the cascading error. This type of error makes the situation even severe as it may have very bad side effects like family stress, cost-consuming, time-consuming and putting more loads on hospital staff.*

Furthermore, the data consumers' answers of the questions, 'Does the importance of this dimension vary along with the task at hand?' and 'Does the importance of this dimension vary along with the class of user at hand?' showed that a dimension is not affected by having different tasks and classes of user. However, all data consumers perceived that some dimensions are more important than others. Moreover, the measures within one dimension are not equally important. This finding was gained through the interviewees' responses of the question, 'from your point of view, do these measuring items have the same weight in assessing the information in terms of "dimension"?'

5.3 Chapter Summary

Questionnaires and semi-structured interviews were analysed to investigate the proper data quality framework for assessing the quality of data in EHR systems in Saudi Arabia. The quantitative and qualitative data confirmed the dimensions proposed in the framework. The proposed 11-dimensional framework was approved as the perceived aspects of data quality. However, there were some issues relating to definitions of some of the dimensions. This confirmation was obtained through both quantitative and qualitative methods.

With regard to data quality items making up the dimensions, most of the items were confirmed as good and relative measures for the respective dimensions, while others were good measures but not relevant to the respective dimensions; they were moved to the appropriate dimension. On the other hand, some measuring items were not sound and were removed from the list. Moreover, the integrity-related dimensions' measures need revision from security experts as the participants involved in this confirmation stage were not confident enough to review these measures.

It is worth noting that, through the interviews, many factors of severity were explicitly and implicitly proposed. These factors would influence the impact of data quality problems on data, therefore need to be taken into consideration during the assessment stage.

Chapter 6 Discussion

This chapter highlights various observations from the literature review and the empirical findings discussed in the previous chapter. They jointly confirm that an appropriate data quality framework could assess the quality of data residing in EHR systems. Moreover, they raise some salient points with regard to the data quality items representing the 11-dimensional data quality framework.

The reviewed literature showed the importance of the 11 dimensions of data quality in EHR: systems, accuracy, consistency, completeness, timeliness, data usability, relevance, provenance, secure access, confidentiality and privacy. The importance of these dimensions was confirmed by the results and findings of the empirical research.

The literature regarding quality problems in the context of databases was reviewed for objective assessment. Subjective assessment-related items were investigated and put to the expert and data consumer groups for review and validation.

6.1 Data quality framework

It was observed in our empirical investigation of the dimensionality of data quality that Saudi EHR stakeholders' perceived data quality framework consisted of 11 dimensions. This observation was drawn from the literature review, the results of the questionnaire and the findings of the interviews, answering the following research question:

RQ1: What data quality determinants are important for EHR stakeholders perceived data quality?

The result of the questionnaire show that the stakeholders of EHR systems agreed on the importance of all the proposed items. This indicates that the dimensions are important to the quality of data consumed by them. However, some additional suggestions were made by three participants: 'availability', 'integrity', 'system security' and 'user friendly' (see Section 5.1). Availability was partly addressed by the dimension of data usability, and clearly needs to be included in the data usability measures. With regard to integrity, it is a concept that captures many dimensions such as accuracy and consistency (Lee et al., 2002; Mandke & Nayar, 1997), therefore this suggestion is disregarded because it is represented by other dimensions in the framework. Regarding system security and user friendliness, the former is within our data quality framework, while user friendliness is unclear and usually associated with systems interfaces, which are beyond the scope of this study.

The findings of the interviews show the highly positive attitudes of interviewees towards these dimensions and their roles in the assessment stage. It was noticeable that the most common terms used were ‘important’, ‘essential’ and ‘no question about it’. However, there were some concerns relating to the definitions associated with some dimensions. These were pointed out in Section 5.2.1.1 and Section 5.2.2.1. The most controversial was the definition associated with completeness, where five of the interviewees claimed it was unclear, especially the second part of the definition, so we adopt the definition of completeness provided in Wang and Strong (1996) as this reflects those aspects that concern the interviewees and is state-based. A team leader said that the definition associated with privacy needed attention. He was not clear about his concerns, and later confirmed that the aspects of consistency needed for consistency measurement were indeed covered, so his concern was disregarded as all the other interviewees agreed on the definition. Regarding concerns over the definition of privacy, a senior DBA called attention to a valid point discussed in Section 5.2.1.1. Therefore, the definition for privacy provided by Rindfleisch (1997) is adopted, as it represents the ideas of the interviewees and covers some aspects addressed by its quality items. Moreover, a medical director, accompanied by two consultants, made a highly important point with regard to the definition associated with secure access. The definition addresses only personal data, while all EHR data are supposed to be protected against any unauthorised access. Hence, the word ‘personal’ was removed. The last observation concerned the definition associated with accuracy. The medical director stated that the definition is difficult for end-users to comprehend due to the use of the term ‘actual values’, and suggested replacing this with ‘documented values’. This is not a valid argument for two reasons. First, documented values are not always correct, and this expression would always make documented values the reference data for any assessment. Secondly, quality measurement of data is usually conducted by a trained team as it comprises procedures and steps that the whole team needs to be aware of, not merely end-users.

The conclusion from our empirical research, supporting our proposition, is that the data quality framework in relation to EHR systems in Saudi Arabia consists of 11 relevant dimensions. These dimensions are shown with their definitions in

Table 6-1.

Table 6-1: Dimensions of data quality framework with their definitions

Dimension	Definition
Accuracy	The extent to which registered data conforms to its actual value.
Consistency	Representation of data values remains the same in multiple data items in multiple locations.
Completeness	The extent to which data are of sufficient breadth, depth, and scope for the task at hand
Timeliness	The state in which data is up to date and its availability is on time.
Interpretabilit y	The degree to which data can be understood.
Usability	The ease with which data can be accessed, used, updated, understood, maintained and managed.
Relevance	The extent to which information is appropriate and useful for the intended task.
Provenance	The source of data, shown and linked to metadata about data.
Secure access	Data being protected against unauthorised access
Confidentiality	The state of information being secret or accessibly restricted under a set of rules that limits the access
Privacy	The right and desire of a person to control the disclosure of personal health information.

6.2 Data Quality Items Making up Dimensions

As discussed in the literature, the quality problems were refactored and mapped against corresponding dimensions. Subsequently, the initial data quality framework and its 65 underlying data quality items were discussed and validated with two groups, comprised of experts and data consumers. See Section 5.2.1.2 Section 5.2.2.2. This Section answered the following sub-research questions:

RQ2.1: what are the measuring items for objective data quality assessment?

RQ2.2: what are the measuring items for subjective data quality assessment?

However, the sub-research question **RQ2.2** is not fully answered as the integrity-related dimensions needed a further reviewing with security experts (see Chapter 1).

6.2.1 Accuracy

Experts and data consumers confirmed that all items associated with accuracy are sound measures and relevant. They found no overlap between the proposed items. With regard to their sufficiency, they agreed on their adequacy for accuracy assessment, although some suggested additional items.

A senior DBA proposed orphan information as a quality problem that needs to be addressed. Orphan information (so-called ‘dangling data’) is caused by a problem known as ‘Referential integrity violation’, already present in items making up consistency assessment. The other two suggestions, data validation and type of dataset for assessment, are not quality problems.

6.2.2 Consistency

In relation to items associated with consistency, there were many issues pointed out by the experts and data consumers. The first issue, suggested by technical experts, was that the item 'Referential integrity violation' belonged to the accuracy measures, not to consistency. This is true, as the root cause of this problem is that wrong data has been entered into the foreign-key field. Thus, this item was moved to the accuracy measures.

With regard to suggestions by two DBAs, the item 'Wrong categorical data' is not rationally accuracy-related as claimed, but is associated with consistency as its value is not considered as an incorrect value but as a user-specified term. However, the two other items 'Contradicting records in single/multi source(s)' and 'Inconsistent spatial data' are indeed likely to be associated with accuracy, as claimed. This is due to the fact that these quality problems are triggered by incorrect data. The claim of 'inconsistent spatial data' being inapplicable might be logical, but cannot be omitted as it received no objection apart from that by a paediatric consultant.

It is worth noting that several professionals expressed concern at having different spellings for one name. They had witnessed many incidents of this quality problem that had ended in serious danger due to inconsistent naming.

6.2.3 Completeness

According to the findings of the interviews from the experts and professionals, they responded positively to completeness-related quality items. They agreed on them being adequate, relevant and comprehensive for completeness assessment.

6.2.4 Timeliness

Interviews, both experts and professionals, responded positively to timeliness-related quality items. However, the two DBAs considered the item 'Outdated reference' ineffective, as such problems can be avoided using RDBMS features. This may be true in an ideal situation, where all data sources enforce such features to allow cascading updates. However, the problem could arise as a result of integration with legacy systems, so it is recommended to keep this item within the timeliness measures.

6.2.5 Interpretability

Interpretability is the most controversial dimension, due to disagreement about its quality items. Many interviewees argued that some of these items address the quality problem relating to consistency. They discussed the problem of the existence of abbreviations and cryptic values, not the different representations, because of their usage, apart from a professional who made the valuable comment that any problem that causes different representations for the same object should be considered a measure of consistency. This is backed by the fact that interpretability is assessed subjectively (Naumann & Rolker, 2000) and can be measured objectively. Therefore, its quality items were moved into the consistency measures and the measures produced by Lee et al. (2002) was adopted.

6.2.6 Data usability

According to the interviews with the two groups, five quality items were approved as good measures for usability with no opposition, and it was suggested that the item 'Is this information easily understood' addresses the interpretability dimension. Besides being suggested by only one interviewee, the widely accepted data quality framework AIMQ in Lee et al. (2002) clearly differentiates between interpretability and comprehensibility when it deals with them as two different dimensions. Thus, this suggestion was disregarded.

Furthermore, the item 'Is this information usable' and its reverse was omitted, as almost all interviewees rejected them since they claimed that these items are too generic and not measurable. The other issue for the interviewees was that the term 'manipulate' is a negative word, so this item was reworded as 'This information is easy to manipulate to meet our needs', as developed in Lee et al. (2002). An expert also added that the update and manipulate items overlap, so we removed 'Is this information easily updatable' as it is less used.

6.2.7 Relevance

According to the findings drawn from the interviews with experts and professionals, most of the interviewees suggested removing the items 'Is this information relevant to the task at hand' and its reverse, as they are generic and would rationally overlap with all other items. The two items were thus removed from the measures.

6.2.8 Provenance

Quality items making up the dimension of provenance were satisfactory and important, according to the interviews conducted with experts and professionals. Neither suggestions nor additional items were proposed.

6.2.9 Secure access

Findings relating to the quality items of secure access show that all proposed items achieved a good level of satisfaction apart from the item ,'Is unauthorised access to this information sufficiently prevented', which was believed to overlap with another item.

There were many suggestions made by interviewees, drawn from their expertise. Measures within this set of measures that need to be included are first, checking data recoverability, and secondly, in the case of remote access, checking whether the remote access policy is enforced.

It is worth noting that some interviewees suggested that this set of measures should be discussed with experts in security to arrive at strong and reliable measures for secure access. This proposition is taken into consideration for future work.

6.2.10 Confidentiality

It was clear from the interviews with the experts and data consumers that they agreed on most of the quality items regarding confidentiality. However, the item 'Is the access to this information granted only to persons who should see it' was removed as redundant, duplicating another item. Moreover, all data consumers rejected the item 'Is the confidentiality of this information achieved' because it was generic and would overlap with all other items as it derived from the term 'confidentiality'. A senior DBA made a valid point when he claimed that vulnerability is linked to the secure access dimension not in this set of measures. This point is valid as vulnerability is a state in which flaws or weakness in systems may be exploited to gain unauthorised access to information (Longley & Shain, 1987), so the item related to vulnerability was moved to that of secure access. Some interviewees suggested that the items associated with this dimension need to be reviewed by security experts.

6.2.11 Privacy

According to findings drawn from the interviews carried out with experts and data consumers, most of the interviewees agreed that PRI3 and PRI4 address the same aspect of privacy as PRI2. Two of the professionals debated whether PRI5 should be removed, while the others believed in its functionality towards privacy assessment. Therefore, this item is subject to more discussion in future work. Apart from these points, they approved the remainders as sound measures for privacy.

To sum up, data quality items were discussed with the expert and data consumer groups to arrive at a valid and reliable instrument with which to assess the quality of EHR data. Most of the feedback was clear, and the researcher reflected upon all but those items associated with secure access, confidentiality and privacy. Within these dimensions, there was no clear pattern of feedback, so it was difficult to make decisions with any certainty. The items relating to these dimensions are to be discussed with security experts to arrive at a solid set of measures for our framework.

Table 6-2: Final version of data quality measures

Accuracy: The extent to which registered data conforms to its actual value.

- Illegal values due to invalid domain range
- Misspellings
- Misfielded values
- Embedded values
- Word transposition
- Wrong reference
- Erroneous entry
- Contradicting records in single/multi source(s)
- Inconsistent spatial data
- Referential integrity violation

Consistency: Representation of data values remains the same in multiple data items in multiple locations.

- Violated attribute dependencies
- Uniqueness violation
- Naming conflicts in multi-source
- Structural conflicts in multi-source
- Wrong categorical data
- Duplicated records in single/multi data source(s)
- Different measure units in single/multi source(s)
- Syntax inconsistency
- Inconsistent name spelling
- Different representations due to use of abbreviation and cryptic values
- Different representations due to use of Alias/nickname
- Different representations due to use of encoding format
- Different representations due to use of special characters

Completeness: The extent to which data are of sufficient breadth, depth, and scope for the task at hand

- Missing data where Null-not-allowed constraint enforced
- Missing data where Null-not-allowed constraint not enforced
- Missing record
- Ambiguous data due to incomplete context
- Semi-empty tuple

Timeliness: The state in which data is up to date and its availability is on time.

- ○ Outdated temporal value
- ○ Outdated reference

Interpretability: The degree to which data can be understood.

- ○ It is easy to interpret what this information means.
- ○ This information is difficult to interpret. (R)
- ○ It is difficult to interpret the coded information. (R)
- ○ This information is easily interpretable.
- ○ The measurement units for this information are clear.

Usability: The ease with which data can be accessed, used, updated, understood, maintained and managed.

- ○ Is this information easily accessible
- ○ Is this information easily retrievable
- ○ Is this information promptly accessible when needed
- ○ Is this information easily understood
- ○ Is the availability of information for the patient treat adequate
- ○ This information is easy to manipulate to meet our needs

Relevance: The extent to which information is appropriate and useful for the intended task.

- ○ Is this information useful to the task at hand
- ○ Is this information applicable to the task at hand
- ○ Is this information appropriate for the task at hand

Provenance: The source of data, shown and linked to metadata about data.

- ○ Is the origin of this information clearly exist
- ○ Is this information owned by known subject
- ○ Is the creation date of this information shown
- ○ Is the update history of this information exist

Secure Access: Data being protected against unauthorised access

- ○ Is the access to this information sufficiently restricted
- ○ Does the access to this information require authentication process
- ○ Is the owner/creator of this information authorised
- ○ In case of system failure, is data safely recoverable
- ○ In case remote access, is remote access policy applied
- ○ Is this information vulnerable

Confidentiality: The state of information being secret or accessibly restricted under a set of rules that limits the access

- ○ Is the confidentiality of this information achieved
- ○ Is this information only accessible by authorised people
- ○ Is the sensitivity of this information clearly declared
- ○ Could this piece of information be disseminated without permission
- ○ In case of sharing, is there a clear consent for this information to be shared

Privacy: The right and desire of a person to control the disclosure of personal health information.

- Is personally identifiable private information being appropriately safeguarded
- Is this personally identifiable information compliant with privacy policy
- Is the access to this personally identifiable information granted only to persons who should see it
- In case of an individual being identified, is there a clear consent from this individual to be identified

6.3 Interpretation of the Assessment Results

Our empirical findings and subsequent discussion with EHR stakeholders resulted in severity factors that would give some measures more weight than others. It was noticed that some factors made quality problems severe, with a deleterious impact on data. This was implicitly observed when discussing quality problems, or observed explicitly though direct questioning. These factors are listed in Table 6-3, answering the following research question:

RQ3: what are the severity factors that make data quality problems more severe?

Table 6-3: Identified severity factors that make data quality problems even worse

Severity factor
Error distribution
Data sensitivity
Data quality measures
Type of data collection tools
Closeness to the correct values
Cascading errors

Finally, it was clearly observed that the importance of a dimension is not affected by having different tasks and classes of user. This finding was gained through the professionals' answers of the questions, 'Does the importance of this dimension vary along with the task at hand?' and 'Does the importance of this dimension vary along with the class of user at hand?' However, the empirical results showed that dimensions are not equally important as some dimensions are believed to be important than others. That would apply to the measures within one dimension as some are perceived more essential than others. This finding was gained through the interviewees' responses of the question, 'from your point of view, do these measuring items have the same weight in assessing the information in terms of "dimension"?'

6.4 Chapter Summary

The results of the questionnaire and the findings drawn from the interviews were examined to answer the three research questions in this chapter, and to refactor the relevant quality problems in the literature. With regard to the first research question and according to the triangulation data, it can be concluded that the 11-dimensional data quality framework shows good validity for data quality assessment in Saudi Arabia.

With regard to the second research question, some of the proposed quality items making up each dimension were approved and confirmed as sound measures, while others did not prompt outright satisfaction. Quality items (measures) associated with secure access, confidentiality and privacy, over which the interviewees engaged in debate, required review with experts in security.

With regard to the third research question, many factors affecting the severity of quality problems were introduced to enhance the result of the data quality assessment. These would influence the impact of data quality problems on the data, therefore need to be taken into consideration during the assessment stage.

Chapter 7 Revision of Integrity-related Dimensions

As discussed in the previous chapter, there were some uncertainties about the integrity-related dimensions and their associated measures amongst IT professionals and data consumers interviewed earlier. The study aims first to confirm the security aspects of data quality in order to keep data secure and safe, and second to discuss and develop measures to measure each of these security aspects. This study was needed to clarify the uncertainty accompanying the outcomes of the EHR stakeholders' interviews.

Following the literature, three integrity-related dimensions were identified as important aspects of data quality: secure access, confidentiality and privacy. Discussions detailed in Sections 6.1 and 6.2 indicate that the data quality dimensions and measuring items needed a further review by security experts. Hence, integrity-related items needed to be reviewed and the opinions and advice of security expert sought. To achieve this task, experts were approached to participate in a group interview as a technique to examine and validate the dimensions of security. This chapter confirmed and answered the sub-research question RQ2.2 and confirmed the measures of integrity-related dimensions. It also contributed to the research question RQ1 as 'secure access dimension' was removed and 'integrity dimension' was added as an outcome of the focus group.

7.1 Focus Group Findings and Results

This Section will briefly describe the findings and results of the focus group conducted with the security experts in healthcare. The focus group was conducted with four experts in order to examine and confirm the potential of the proposed integrity-related dimensions and their measuring items (measures), and the possibility of enhancing this set of measures by tackling more aspects of data quality problems not covered in the proposed measures. After analysing the group interview with the security experts, the output was themed into the proposed integrity-related dimensions.

7.1.1 Secure access

Expert B suggested that 'secure access' is not the proper term commonly used in the context of security, and that it is 'access control'. This proposal was backed and supported by the remainder of the focus group candidates. Subsequently, Expert A argued that access control should not be a dimension of its own, but rather falls under confidentiality dimension. The following quotes are provided in support of these arguments:

Expert B: *Secure access is not a common term used in the security context. Instead, access control is appropriate and commonly used.*

Expert A *I agree that access control is a proper term to describe it, and means to ensure access to an asset is authorized, based on security requirements. However, access control is a technique or practice, not a security concept. It is one of the ways that ensure confidentiality.*

Expert C *I completely agree with Expert A. Access control is a mechanism to ensure authorization and confidentiality for patient records, but not a dimension, like confidentiality.*

7.1.1.1 Data quality items

Although the secure access dimension was not approved to be a security dimension by all, its measuring items were discussed to explore the possibility of using them to measure another dimension. Table 7-1 presents the outcomes of the discussion detailed in Chapter 1 regarding the measuring items of secure access dimension.

Table 7-1: Measures of secure access

Item code	Data quality item
ACC1	Is the access to this information sufficiently restricted
ACC2	Does the access to this information require authentication process
ACC3	Is the owner/creator of this information authorised
ACC4	In case of system failure, is data safely recoverable
ACC5	In case remote access, is remote access policy applied
ACC6	Is this information vulnerable

As shown in Table 7-2, items ACC1, ACC2 and ACC5 are believed to be good measures of the confidentiality dimension, suggested by Experts B and D, and agreed by the remainder.

Table 7-2: Security experts' findings for secure access measures

Item code	Suggested dimension	Expert A	Expert B	Expert C	Expert D
ACC1	Confidentiality	Agree	Suggest	Agree	Agree
ACC2	Confidentiality	Agree	Suggest	Agree	Agree
ACC3	Integrity*	Agree	Agree	Suggest	Agree

ACC4	Integrity*	Suggest	Agree	Agree	Agree
ACC5	Confidentiality	Agree	Agree	Agree	Suggest
ACC6	Not approved	Disagree	Disagree	Disagree	Disagree

* Dimension not introduced yet

Item ACC3 was suggested as a good measure for integrity, though there was no such dimension in the proposed data quality framework, to be discussed later. Moreover, all experts agreed on the importance of item ACC4. It measures the ability to restore the affected data to its correct state, in case of system failure. Lastly, ACC6 was not approved as a measure for any dimension in hand.

7.1.2 Confidentiality

With regard to confidentiality as a security dimension, there was general agreement on the importance of this dimension as an essential aspect of security, especially in healthcare field. However, the experts raised some issues regarding the definition of confidentiality. Their recommendations were to use the definition provided in ISO 27001. The following quotes demonstrate these arguments:

Expert C: *Confidentiality is an essential aspect of security. Medical information must be confidential, and access is granted to authorized subjects. I am not sure about the definition provided, but I prefer the definition provided by international standards, like ISO.*

Expert D: *There is no doubt about the importance of confidentiality. But the definition needs attention. I would suggest ISO27000, instead.*

Expert A: *I agree with Expert D, as the confidentiality definition given by ISO 27001 is better worded.*

7.1.2.1 Data quality items

Six items were the output of the focus group regarding the measuring items of confidentially. Only CONF3 and CONF5 from the initial set were confirmed and approved (see Table 7-3). Expert D proposed one measuring item, and this item was approved by the remainder.

Table 7-3: Confidentiality measures

Item code	Data quality item
CONF1	Is the confidentiality of this information achieved
CONF2	Is this information only accessible by authorised people
CONF3	Is the sensitivity of this information clearly declared
CONF4	Could this piece of information be disseminated without permission
CONF5	In case of sharing, is there a clear consent for this information to be shared

The proposed item during the session measures whether the access to medical information is secure. It was articulated as "access to the medical information is secure". Nevertheless, all the experts apart from Expert B claimed that CONF1 was too generic, while CONF2 would overlap with the measuring items brought from the disapproved dimension (secure access). Furthermore, CONF4 was seen as being impractical.

Table 7-4: Security experts' findings for confidentiality measures

Item code	Expert A			Expert B			Expert C			Expert D		
	C	I	O	C	I	O	C	I	O	C	I	O
CONF1		✓		✓				✓			✓	
CONF2			✓			✓			✓			✓
CONF3	✓			✓			✓			✓		
CONF4		✓			✓			✓			✓	
CONF5	✓			✓			✓			✓		

C: Confirmed item | I: Inappropriate item | O: Overlapping item

7.1.3 Privacy

Experts considered privacy as an essential facet of data quality, especially in the healthcare field. In term of its definition, there was no question about its wording and meaning. The following quotes support these arguments:

Expert A: *Privacy is very crucial in EHR systems. And its definition is sound for me.*

Expert B: *Privacy is a big concern in healthcare settings. Protecting patients' privacy is always priority. For me, its definition reflects the meaning of privacy.*

7.1.3.1 Data quality items

There was an initial set of four measuring items of privacy brought in for discussion, as shown in Table 7-5. Three of the initial set of items were approved and confirmed as good measuring items. Two of the three, claimed by Experts B, C and D, needed rewording for the sake of clarity. They came to the conclusion that PRI1 and PRI2 should be reworded as follows:

PRI1: "patient's privacy is properly protected"

PRI2: "privacy policy is enforced for patient data access"

Table 7-5: Privacy measures

Item code	Data quality item
PRI1	Is personally identifiable private information being appropriately safeguarded
PRI2	Is this personally identifiable information compliant with privacy policy
PRI3	Is the access to this personally identifiable information granted only to persons who should see it
PRI4	In case of an individual being identified, is there a clear consent from this individual to be identified

As displayed in Table 7-6, all experts argued that item PRI3 overlapped with PRI1, as the two items measure the same thing. Moreover, Expert A emphasized the importance of the anonymity approach in preserving the patients' privacy, so he suggested adding an item that would measure, in case of data release, whether the anonymized data met with privacy protection. This suggestion gained the approval of the remainder.

Table 7-6: Security experts' findings for privacy measures

Item code	Expert A			Expert B			Expert C			Expert D		
	C	I	O	C	I	O	C	I	O	C	I	O
PRI1	✓			✓			✓			✓		
PRI2	✓			✓			✓			✓		
PRI3			✓			✓			✓			✓
PRI4	✓			✓			✓			✓		

C: Confirmed item | I: Inappropriate item | O: Overlapping item

7.1.4 Integrity

As mentioned in Section 7.1.1, during the discussion of the measuring items associated with 'secure access', experts emphasized the need to include the security dimension 'integrity' in the proposed data quality framework. Their rationale was that integrity maintains the consistency, accuracy and trustworthiness of data. All four security experts firmly suggested the inclusion of integrity. The following quotes demonstrate their arguments:

Expert D: *Integrity is an important pillar of the CIA triad security model.*

Expert A: *I was surprised that integrity was not included in your data quality framework. It is so important!*

Expert C: *Integrity is a must-have dimension. It is the dimension for preserving data from unauthorized changes.*

They stressed the role of integrity dimension in assessing data quality of EHR systems. It ensures that data, either confidential or non-confidential, are protected from unauthorized changes. With regard to its definition, ISO 27001 was recommended by Expert C.

7.1.4.1 Data quality items

As discussed in Section 7.1.1, ACC3 and ACC4 were proposed by the experts as good measuring items of integrity. Expert D introduced a measure that would make sure that any transactions would comply with authorization policy. This suggestion was backed Experts B and C. Moreover, Expert B contributed to the measuring items by highlighting the need for proper enforcement of security privileges; that is, whether security privileges are properly applied (INT1).

Table 7-7: Integrity measures identified by the security experts

Item code	Data quality item	Suggested by
INT1	Security privileges are properly applied	Expert B
INT2	The owner/creator of this information is authorised	Secure access*
INT3	Transactions (add/delete/modify) made to this information are in line with authorization policy	Expert B
INT4	In case of system failure, data is safely recoverable	Secure access*

* Disapproved dimension

7.2 Discussion of Security-related Dimensions and Associated Measuring Items

This Section discusses the results presented in Section 7.1 to identify the security aspects of data quality in the context of EHR systems, including confidentiality, integrity and privacy. It also discusses the measuring items potentially used to assess the data quality of EHR systems from the security perspective.

7.2.1 Secure access and associated measuring items

Security aspects of data quality in the healthcare sector are crucial, and it is necessary to protect patients' data from the growing security threats due to shifting from paper-based to electronic systems, as well as the advances in ICT, and maintain this protection (Farzandipour et al., 2010). It was observed in our empirical investigation of security facets of data quality that Saudi EHR stakeholders' perceived security-related data quality framework consisted of three dimensions: confidentiality, integrity and privacy.

Initially, secure access was introduced into the proposed framework, but the findings of the interviews of IT professionals in EHR systems discussed in Section 6.1 showed that there was no consensus on the dimension itself and its associated measuring items. Subsequently, the findings of the focus group with security experts in the Saudi Arabia health sector detailed in Section 7.1.1 revealed that ‘secure access’ or ‘access control’ (to word it better) is a security control, not a quality aspect of security. It is a mechanism to ensure the integrity and confidentiality of personal health information (British Standards Institute, 2016). Lee et al. (2002) perceived ‘secure access’ as a measuring item in their developed instrument, under the security dimension used to assess data quality from a security angle.

Therefore, this dimension was removed from the security-related data quality framework, and its measuring items were refactored to see if they potentially fitted another security dimension.

7.2.2 Confidentiality and associated measuring items

Confidentiality is one of the fundamental security goals that are essential for EHR systems. There is a strong need for this dimension, as EHR systems contain highly sensitive information about patients (Haas et al., 2011). As detailed in Sections 6.1 and 7.1.2, confidentiality was perceived as an essential constituent of a security-related data quality framework.

However, there were some issues with the definition initially proposed and its implications. The definition provided by ISO 27799:2016 is better worded and comprehensive. This International Standard defines guidelines to support and facilitate the implementation in health informatics of ISO/IEC 27002.

As detailed in Section 7.1.2 regarding measuring items of confidentiality, five measuring items were approved for confidentiality. Two items were approved and accepted as measuring items of confidentiality from the initial set of items. Also, three measuring items were borrowed from the disallowed dimension, and one was introduced by one expert and approved by the rest.

Table 7-8 shows the final version of the measuring items, as well as the definition of confidentiality.

Table 7-8: Final version of confidentiality measures

Confidentiality	Property that information is not made available or disclosed to unauthorized individuals, entities, or processes	
Item#	Data quality item	Former code
1	Access to this information is sufficiently restricted	ACC1
2	Access to this information requires authentication process	ACC2
3	In case remote access, remote access policy is applied	ACC5
4	Access to the medical information is secure	
5	The sensitivity of this information is clearly declared	CONF3
6	In case of sharing, is there a clear consent for this information to be shared	CONF5

7.2.3 Privacy and associated measuring items

Privacy is a crucial security requirement that EHR systems need to meet in response to patients' concerns over their data. It was perceived by interviewees, as detailed in Section 6.1 and Section 7.1.3, to be an essential security component of data quality. The definition provided for privacy was sound to experts and approved by them.

Regarding measuring items, all were believed to be good measures, apart from PRI3. This overlapped with PRI1. Moreover, an expert highlighted an aspect of privacy that needed to be addressed. This aspect concerned the patient anonymity approach which uses coded patient identifiers (pseudonyms). This approach is usually utilized in the healthcare systems to anonymize at least some parts of the care process (Rindfleisch, 1997), and this method need to be assessed against privacy policy. Table 7-9 displays the final version of the measuring items, as well as the definition of privacy.

Table 7-9: Final version of privacy measures

Privacy	The right and desire of a person to control the disclosure of personal health information	
Item#	Data quality item	Former code
1	Patient's privacy is properly protected	PRI1
2	privacy policy is enforced for patient data access	PRI2
3	In case of data release, anonymized data meets privacy protection	-
4	In case of an individual being identified, is there a clear consent from this individual to be identified	PRI4

7.2.4 Integrity and associated measuring items

The findings of the focus group of security experts, as discussed in Section 7.1.4 and the literature (Gritzalis & Lambrinoudakis 2004; Orfanidis et al. 2004; Lee et al. 2002; Häyrinen et al. 2008), confirmed the importance of integrity as a dimension in assessing data quality of her systems. It is one of the fundamental security objectives. Protecting the integrity of health information ensures patient safety (Fernández-Alemán et al., 2013).

The ISO 27799 definition of integrity was borrowed, as suggested in the findings and results of qualitative data presented in Section 7.1.4. This is because the ISO 27799 standard is meant for managing information security in healthcare organizations.

Table 7-10: Final version of privacy measures

Integrity	Property that data has not been altered or destroyed in an unauthorized manner	
Item#	**Data quality item**	**Former code**
1	Security privileges are properly applied	-
2	The owner/creator of this information is authorized	ACC3
3	Transactions (add/delete/modify) made to this information are in line with authorization policy	-
4	In case of system failure, data is safely recoverable	ACC4

As displayed in Table 7-10, four items were developed to assess the integrity of EHR data. Two of them were introduced during the focus group session, while other two were borrowed from the initial set of items of the excluded security dimension.

7.2.5 Security-related dimensions

According to the findings from the interviews with the security experts, the security-related dimensions are confidentiality, integrity and privacy. 'Secure access' was disregarded and removed for the explanations provided in 7.2.1. The category's name for this group is changed into security-related dimensions. This is, first, to avoid the misconception as the proposed framework has an emerging dimension, integrity. Second, the integrity was introduced in (Tupek, 2006) as one of the attributes the information products must possess (Section 3.1.2). It 'refers to the security or protection of information from unauthorized access or revision'. This implies that 'security' is another synonym of 'integrity'.

7.3 Chapter summary

As discussed in previous chapter, some issues linked with the measures associated with security-related dimensions have arisen during the confirmation stage. Therefore, these issues needed further discussion and reviewing with security experts in order to revise the proposed framework and its measures. So in this chapter, these issues were brought to the security experts, and revised in order to fully cover the security aspects of data quality. After confirming all components of the proposed data quality framework, the next step is to conduct a real case study to examine the practicality of the developed instrument (the data quality framework and its associated measures).

Chapter 8 Case Study

This chapter focuses on the case study and the experimental results. The case study is used to demonstrate our data quality framework and its data quality (DQ) instrument, to understand the usefulness of the DQ instrument and to examine its practicality and applicability. It presents the final version of the data quality assessment instrument. It also presents the data analysis of a large-scale hospital and its quality scores in each dimension. Subsequently, the feedback about the DQ instrument was collected and used as a practicality test. The sub-research question **RQ2.3** to be answered in this chapter was, *How well is the functionality and practicality of the proposed instrument?*

8.1 Final version of the proposed instrument

The development of the proposed instrument went through many stages. After reviewing the literature of the data quality frameworks in healthcare field, eleven data quality dimensions were extracted and proposed, and these eleven dimensions were mapped and grouped into three categories (see Figure 3.4 in Section 3.1.3). The measures associated with each of these eleven dimensions were introduced in Section 3.2. After that, the questionnaire and the interviews jointly confirm that the proposed data quality framework (DQ dimensions and their associated measures) could assess the quality of data residing in EHR systems (see Table 6-2). Subsequently, the security-related dimensions and measures were reviewed with four security experts to reach a valid and reliable set of security dimensions and measures. Table 8-1 presents the final version of the proposed instrument with which the EHR data of a large hospital was assessed. This case study was conducted to examine the practicality of the proposed instrument.

The instrument considered both objective and subjective assessment approaches. Four dimensions are objectively assessed – accuracy, consistency, completeness and timelines. On the other hand, the remainder of the dimensions are subjectively assessed. Some templates used during the assessment stage of the case study are provided in APPENDIX E.

Table 8-1: Final version of the proposed instrument

Dimension	Measures	Aggregation method
Accuracy	○ Illegal values due to invalid domain range ○ Misspellings ○ Misfielded values ○ Embedded values ○ Word transposition ○ Wrong reference ○ Erroneous entry ○ Contradicting records in single/multi source(s) ○ Inconsistent spatial data ○ Referential integrity violation	Dimension quality= $\boldsymbol{avg}\left(\sum_{i=1}^{n}\frac{\boldsymbol{no.of\ data\ units\ free\ of\ M_i}}{\boldsymbol{total\ no.of\ data\ units\ assessed}}\right)$ where M_i is the ith measure within a dimension, and n is the number of measures associated with the dimension.
Consistency	○ Violated attribute dependencies ○ Uniqueness violation ○ Naming conflicts in multi-source ○ Structural conflicts in multi-source ○ Wrong categorical data ○ Duplicated records in single/multi data source(s) ○ Different measure units in single/multi source(s) ○ Syntax inconsistency ○ Inconsistent name spelling ○ Different representations due to use of abbreviation and cryptic values ○ Different representations due to use of Alias/nickname ○ Different representations due to use of encoding format ○ Different representations due to use of special characters	
Completeness	○ Missing data where Null-not-allowed constraint enforced ○ Missing data where Null-not-allowed constraint not enforced ○ Missing record ○ Ambiguous data due to incomplete context ○ Semi-empty tuple	
Timeliness	○ Outdated temporal value ○ Outdated reference	
Interpretability	○ It is easy to interpret what this information means. ○ This information is difficult to interpret. (R) ○ It is difficult to interpret the coded information. (R) ○ This information is easily interpretable. ○ The measurement units for this information are clear.	Dimension quality= $avg(\sum_i^n avg\left(\sum_{j=1}^{p} S_{ij}\right)$ where n is the number of measures associated with a dimension, and p is the number of participants involved in the assessment process. S_{ij} is the value of the assessment of the ith measure and jth participant.
Usability	○ Is this information easily accessible ○ Is this information easily retrievable ○ Is this information promptly accessible when needed ○ Is this information easily understood ○ Is the availability of information for the patient treat adequate ○ This information is easy to manipulate to meet our needs	
Relevance	○ Is this information useful to the task at hand ○ Is this information applicable to the task at hand ○ Is this information appropriate for the task at hand	
Provenance	○ Is the origin of this information clearly exist ○ Is this information owned by known subject ○ Is the creation date of this information shown ○ Is the update history of this information exist	
Confidentiality	○ Access to this information is sufficiently restricted ○ Access to this information requires authentication process ○ In case remote access, remote access policy is applied ○ Access to the medical information is secure ○ The sensitivity of this information is clearly declared ○ In case of sharing, is there a clear consent for this information to be shared	
Integrity	○ Security privileges are properly applied ○ The owner/creator of this information is authorized ○ Transactions (add/delete/modify) made to this information are in line with authorization policy ○ In case of system failure, data is safely recoverable	
Privacy	○ Patient's privacy is properly protected ○ Privacy policy is enforced for patient data access ○ In case of data release, anonymized data meets privacy protection ○ In case of an individual being identified, is there a clear consent from this individual to be identified	

8.2 EHR Assessment Process and Results

The proposed instrument considers both objective and subjective aspects of the data assessment. Figure 8.1 illustrates the data quality assessment framework utilized to assess the quality of the data populating their system. The objective assessment process covers all the objectivity-relates dimensions. These include accuracy, consistency, completeness and timeliness. It measures the retrieved sample data against some measures provided in the DQ instrument. By contrast, the subjective assessment process concerns live data observed on the EHR systems. It includes utility-related and security-related dimensions.

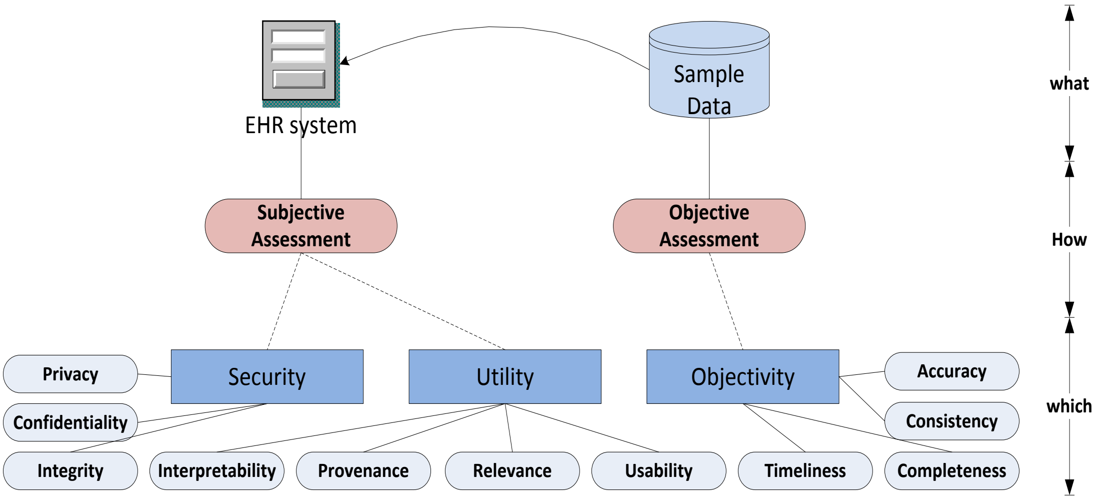


Figure 8.1 Data quality assessment framework

Figure 8.2 shows the result of the assessment process. Lab and pharmacy systems have almost the same quality score for all dimensions. They scored reasonably well, compared to the SOAP system. It was observed that these two systems clearly have enough attention from the top management of the hospital. All processes and procedures are coded, and have quality constraints before sending or releasing any data. Besides, data entry errors were minimized by also coding all pharmacy-related items, diseases and symptoms. These procedures led to better results for the two systems than the SOAP system.

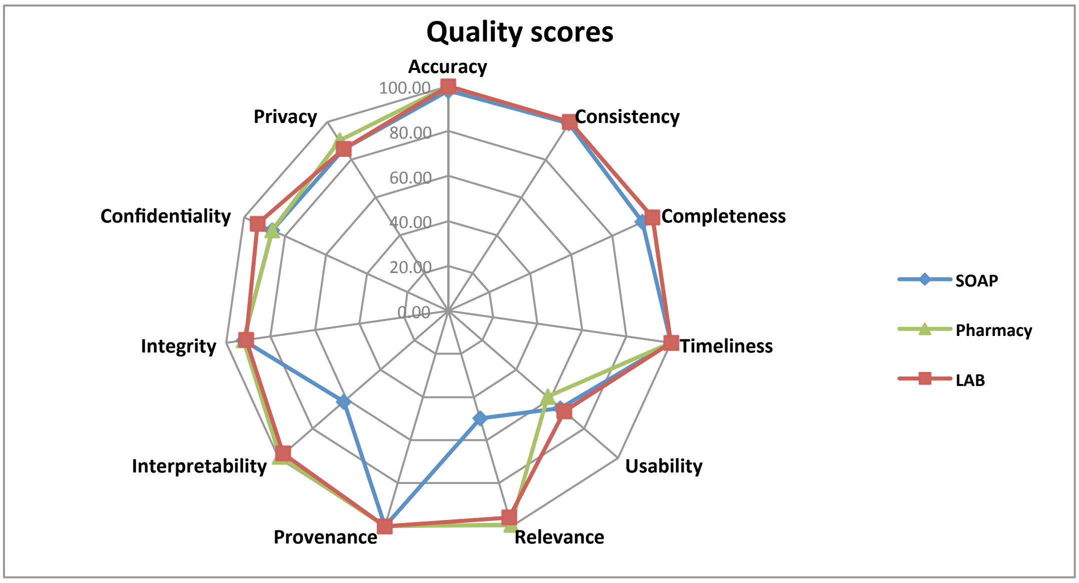


Figure 8.2: Quality scores for the three systems

On the other hand, the SOAP system reported a relatively low score in data quality compared to the other systems. It showed a poor quality of data when it came to relevance and interpretability dimensions. The data quality constraints and precautions for this system need reconsideration.

In general, the EHR system scored a good level of quality in objectivity-related dimensions using the proposed taxonomy. This new approach highlighted some issues with regard to usability. The system also needs attention with regard to the security aspects of data quality, especially in the privacy dimension.

8.2.1 Assessment results of objectivity dimensions

Lab and pharmacy systems score almost 100% in all objectivity-related dimensions, as shown in Figure 8.3; that is, accuracy, consistency, completeness and timeliness. Yet the SOAP system achieved less good quality for completeness, while scoring over 98% in accuracy, consistency and timeliness; these are discussed further below.

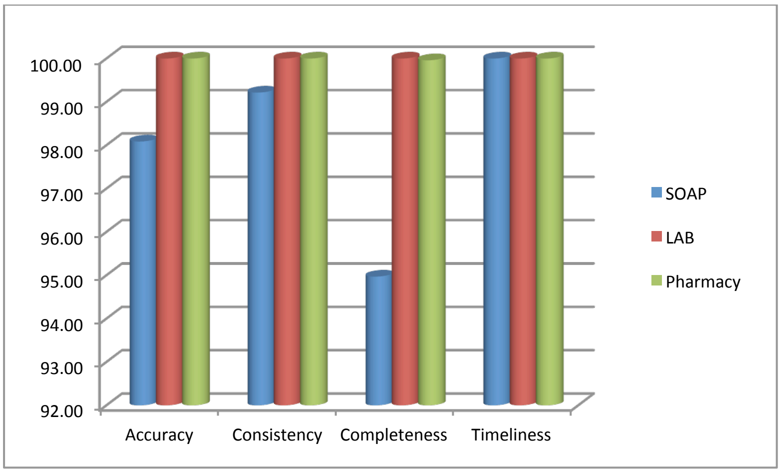


Figure 8.3: Assessment scores of the objectivity-related dimensions

8.2.1.1 Accuracy measures

The accuracy consists of ten measures. ‘Inconsistent spatial data’ was excluded, due to being inapplicable to the dataset of the case study after reviewing the data with the quality assessment team. As mentioned earlier, Lab and pharmacy systems have outstanding results in terms of the accuracy-related quality of the dataset. However, ‘misspellings’, ‘illegal values due to invalid domain range’ and ‘contradicting records in single/multi-source(s)’ measures affected the accuracy of the patients’ data populating the SOAP system.

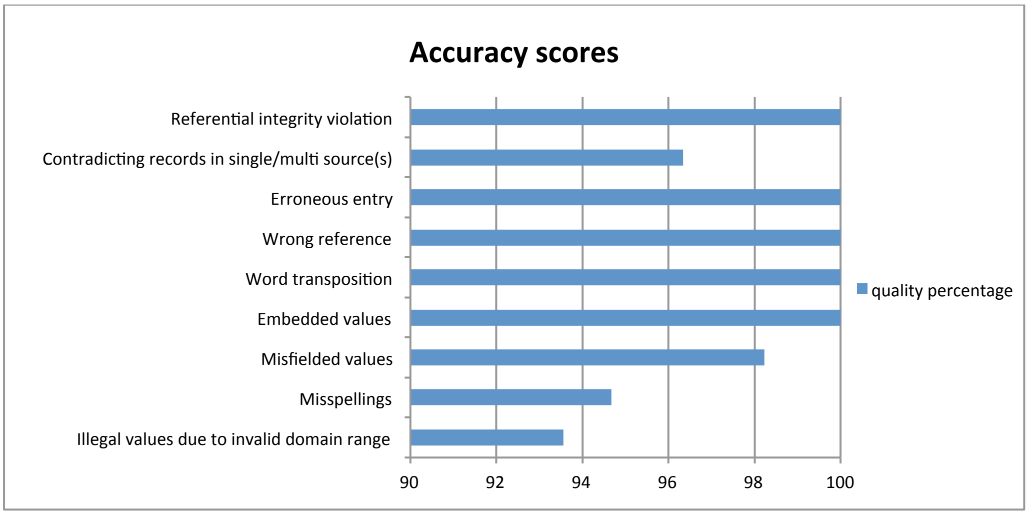


Figure 8.4: Accuracy results for SOAP system

8.2.1.2 Completeness measures

The completeness dimension was assessed through five measures. As illustrated in Figure 8.5, some quality issues emerged in SOAP's quality. The 'semi-empty tuple' measure compromised the quality of the dataset, leading to scoring 80% in the SOAP system.

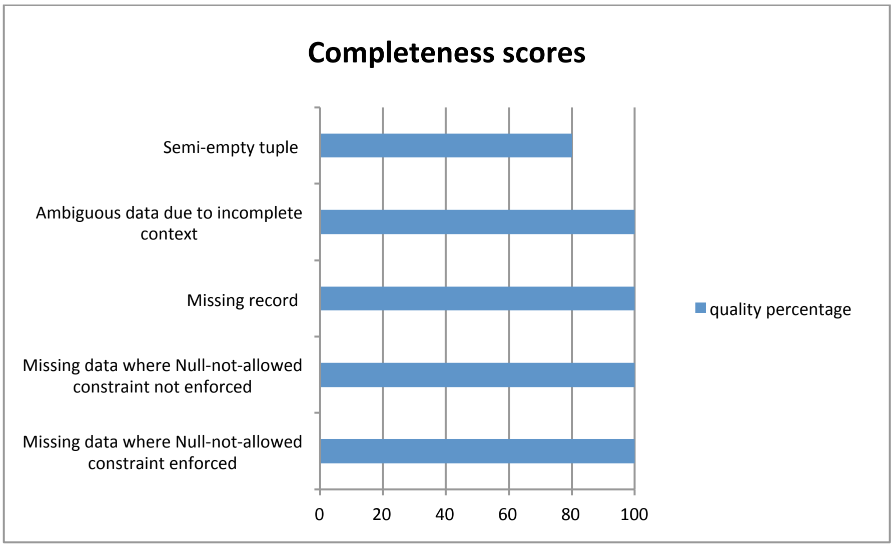


Figure 8.5: Completeness results for SOAP system

The rest of the measures showed the high quality of the dataset, apart from a few issues with an incomplete context, due to some ambiguous data. Regarding the pharmacy system, some minor quality problems arose because null-not-allowed constraint was not enforced, which caused some needed missing data. No quality issues were reported in Lab systems.

8.2.1.3 Consistency measures

Twelve measures were utilized to assess the consistency dimension, discounting 'different representations due to use of alias/nickname' as it was not applicable to the context. No quality issues were reported in pharmacy and Lab systems. As shown in Figure 8.6, however, there were quality issues over consistency in the SOAP system.

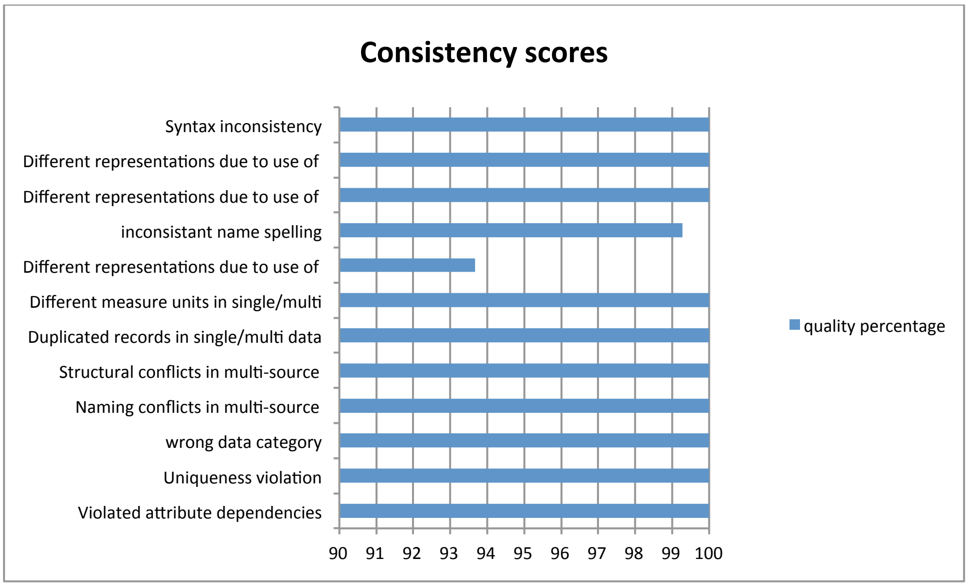


Figure 8.6: Consistency results for SOAP system

The occurrence of different representations due to the use of abbreviation slightly depressed the level of consistency quality. This quality problem was observed in 57 incidences in all datasets. Moreover, it was reported that some family names have different spellings, causing inconsistency in the dataset. This was captured through the 'inconsistent name spelling' measure.

8.2.1.4 Timeliness measures

The results show that there is no concern about the quality of timeliness within the dataset used in the case study. However, it was observed that some timeliness-related quality issues with volatility were not captured by the utilized timeliness measures. It is a measure of information validity over a given time. For instance, some lab requests were labelled as urgent, which means that they need to be processed and the result sent back within two hours. So, if it was received in three hours, it may have up-to-date values but does not meet the consumer's requirements.

8.2.2 Assessment Results of Utility dimensions

As shown in Figure 8.7, the results show that there are no quality issues with the provenance of the dataset sampled from the three systems. However, they show that the usability dimension is of poor quality. Moreover, they indicate that the SOAP system scored a low level of quality in the relevance and interpretability dimensions.

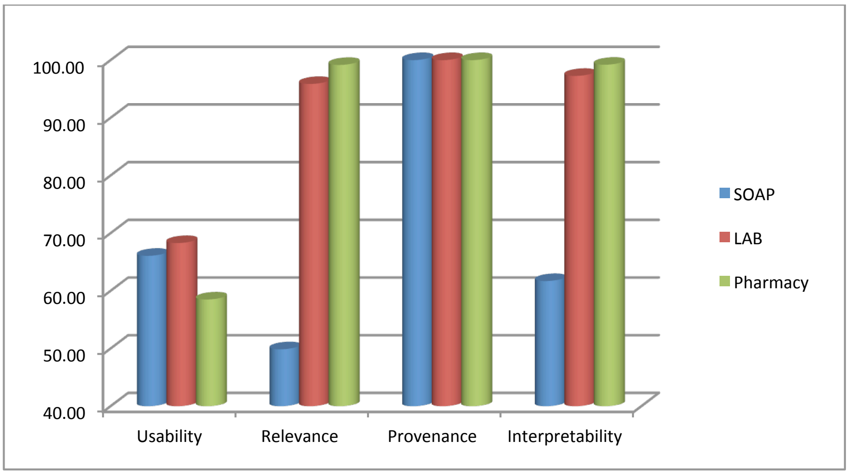


Figure 8.7: Assessment scores of the Utility-related dimensions

8.2.2.1 Usability measures

The result of the usability dimension shows poor quality in the usability of the three systems. As shown in Figure 8.8, the systems' stakeholders show no satisfaction with usability of the three systems for any measuring items. It is clear that the pharmacy system scores lower in almost all measures than the others.

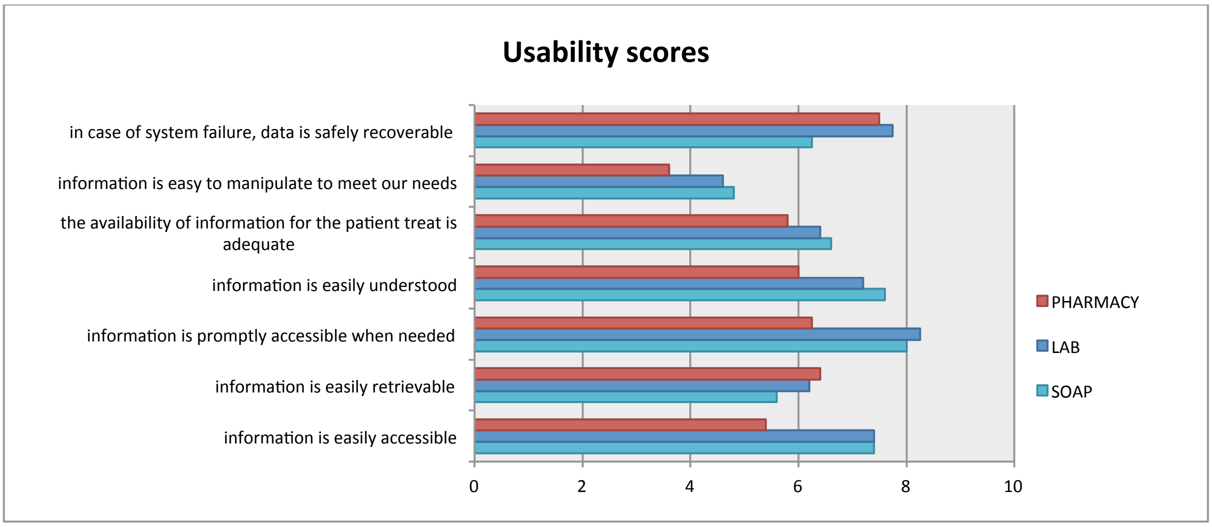


Figure 8.8: Results of the usability measures

Figure 8.8 also shows that the level of data manipulation given to the users is not satisfactory and its corresponding measure scored the lowest result. It also indicates that the facility of information being promptly accessible when needed is relatively good in lab and SOAP systems, and poor in the pharmacy system.

8.2.2.2 Relevance measures

Pharmacy and lab systems show a very high quality in terms of information relevance, as shown in Figure 8.7. The stakeholders show no quality issues during the quality assessment process with the two systems. However, the SOAP system was poor in quality, as illustrated in Figure 8.9. Users of SOAP expressed their concern about the appropriateness, applicability and usefulness of the retrieved medical information.

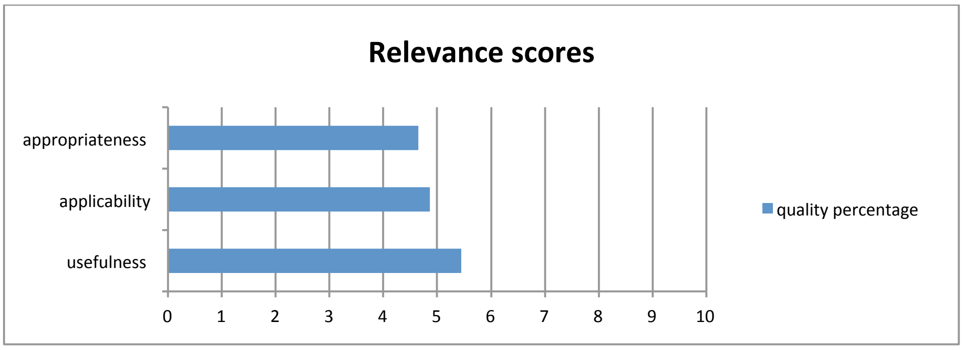


Figure 8.9: Relevance results for SOAP system

8.2.2.3 Provenance measures

The provenance dimension achieved 100% in all measures of all systems. This indicates that the origin of information exists, and is owned by a known subject. The creation date is also shown and there is a history of information.

8.2.2.4 Interpretability measures

The interpretability dimension recorded a high score in quality in both the pharmacy and lab systems, achieving 99.2% and 97.2% respectively. However, the SOAP system scored 61.7%, indicating that some information were not easily interpreted and that a considerable amount of coded information was not clear and difficult to interpret.

8.2.3 Assessment Results of Security dimensions

As displayed in Figure 8.10, the results show that the quality level of security is relatively acceptable. The result of the integrity-related dimensions assessment was over 91% for all systems, while the privacy assessment results ranged from 85% to 90%. With regard to confidentiality dimension, SOAP and pharmacy both recorded 86.1%, whereas the lab system had in a good level of confidentiality, achieving 93.3%.

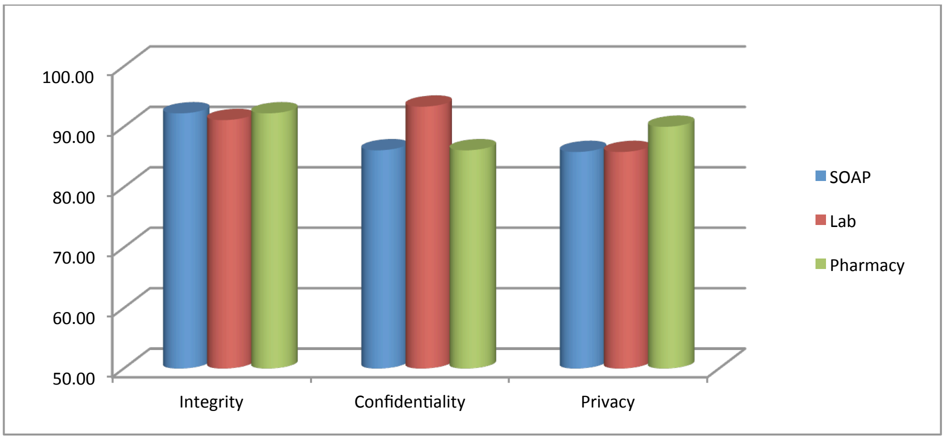


Figure 8.10: Assessment scores of the security-related dimensions

8.2.3.1 Integrity measures

The enforcement of the security privileges in the three systems is acceptable, as the corresponding measure to this issue scored 9.2 for each system. Moreover, SOAP and pharmacy are in a satisfying level of quality in terms of applying the authorisation process achieving 9.3 out of 10, while the lab system recorded a lower quality score.

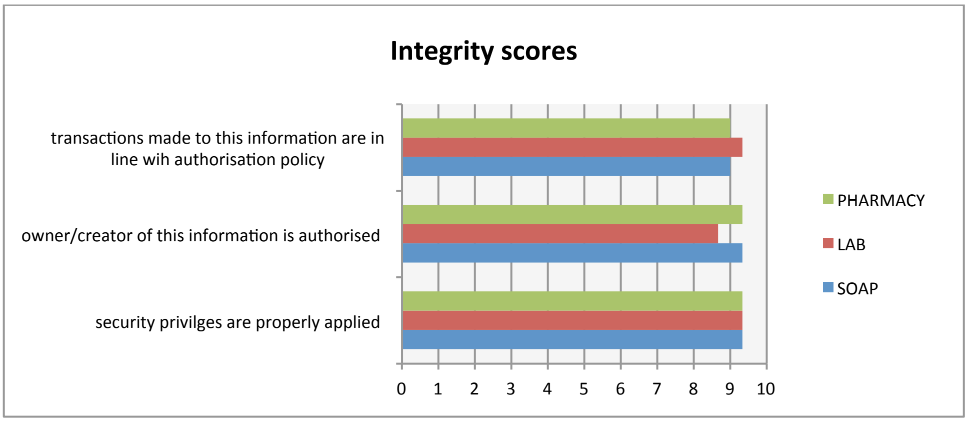


Figure 8.11: Results of the integrity measures

8.2.3.2 Confidentiality measures

As shown in Figure 8.12, the quality assessment team positively confirmed the existence of consent in the case of information sharing. This is clear, as they rated it at over 9.5 for the three systems. Regarding the declaration for sensitive information, the related measure is reasonably good, scoring over 9 for each system. Furthermore, the enforcement of authentication process gains a good level of quality, achieving 9.3 out of 10 for each system.

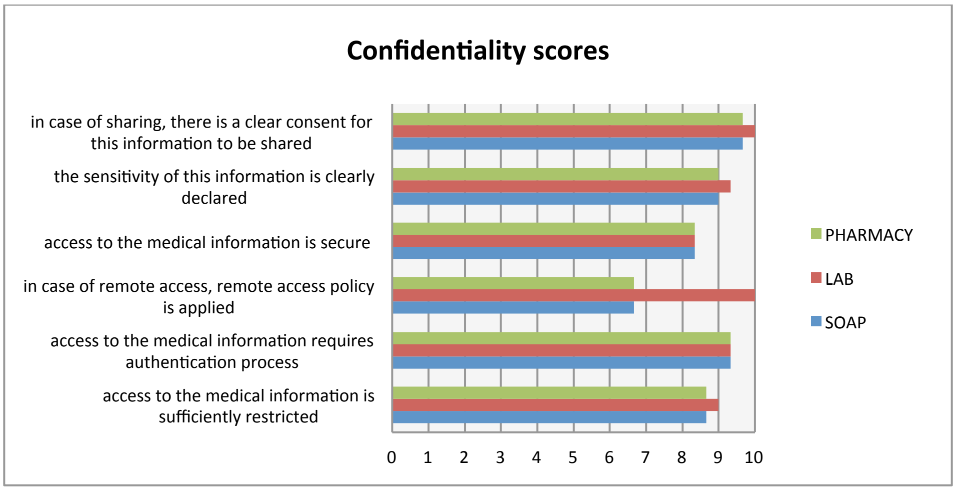


Figure 8.12: Results of the confidentiality measures

However, the result of the secure access-related measure shows that the stakeholders had some concerns about the secureness of the access to the medical information residing in all systems, scoring 8.3 out of 10. Besides, the sufficient access restriction measure scored 8.6 out of 10 for the SOAP and pharmacy systems, and 9 for the lab system. It is worth mentioning that the enforcement of the remote access policy triggered a worrying quality issue for the SOAP and pharmacy systems, recording 6.6, while the lab system scored 10.

8.2.3.3 Privacy measures

As shown in Figure 8.13, patients' privacy is protected and maintained, and the privacy policy was soundly maintained and enforced for the three systems. However, the lab and SOAP systems have some quality issues with regard to the consent form, scoring 8.3 and 7.6 respectively. Furthermore, the release of the anonymized data was not properly maintained in all systems, as the respective measures show that this practice did not meet the privacy protections.

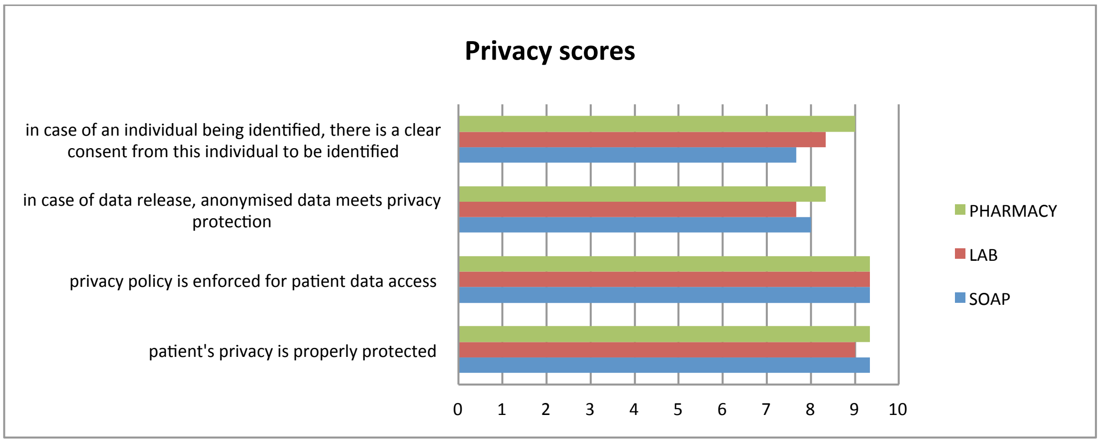


Figure 8.13: Result of the privacy measures

8.3 Practicality of the DQ Instrument for the Case Study

The next step was to understand the practicality and usefulness of the proposed DQ instrument. This process was achieved through a questionnaire distributed to the quality assessment team and interviews with two senior managers. First, 10 members of the quality assessment team were asked to evaluate the practicality of the DQ instrument through a questionnaire developed by the researcher. The questionnaire examined 'the ease of the use', 'its perceived usefulness', 'user satisfaction' and 'the perception of congruence between expectation of the use and its actual performance' of the proposed approach. Subsequently, semi-structured interviews with two senior managers were conducted to discuss the results of the quality assessment process along with the usefulness of the proposed DQ instrument.

8.3.1.1 Reliability of the survey items

The questionnaire was distributed and collected from 10 out of 12, quality assessment team members representing 83% of the total population. No concerns were, expressed with regard to the questions. To establish the reliability of the survey question items for each scale, and to assess how far each set of question items produced consistent results, construct reliability was tested using the Cronbach alpha. Cronbach alpha was used to determine the internal consistency between a set of items measuring the same variable (Cronbach, 1951).

Table 8-2: Cronbach's alpha value of each set of items for each survey construct

Perceived ease of use	4	• Learning to operate **Data Quality Assessment tool** is easy for me • I find it easy to get **Data Quality Assessment tool** to do what I want • It is easy to become skilful at using **Data Quality Assessment tool** • Overall, I find **Data Quality Assessment tool** easy to use	0.80
Satisfaction	4	• I am satisfied about the quality results I got after using the tool • I am pleased for the overall quality of our data after using the tool • I am content with the experience of using the tool • How would you rate your overall satisfaction with us?	0.75
Perceived usefulness	4	• Using this tool helps assess the quality of our system data. • Using this tool increases my productivity in assessing and measuring the quality of our medical data. • Using this tool enhances my effectiveness in managing and assessing the quality of our medical data. • Overall, this tool is useful in assessing and assuring the quality of our medical data.	0.71
Confirmation	2	• My experience with using this tool was better than what I expected. • The service level provided by this tool was better than I expected.	0.93

As shown in Table 8-2, values that range from 0.93 to 0.71 exceed the threshold value of 0.70. This indicates that the measures of each survey construct are reliable. The questionnaire was to understand the practicality of the DQ instrument in assessing the data quality from the EHR stakeholders' perspectives.

8.3.1.2 Survey Data Analysis

Ten of the quality assessment team was involved in this questionnaire-based evaluation process, since two consultants went on holiday after the completion of the data quality assessment, hence they did not complete the questionnaire.

The questionnaire data was analysed using SPSS software to examine the perception of the quality assessment team towards the DQ instrument. Table 8-3 presents the output for the one-sample t-test conducted to determine whether the mean rating for each question was significantly different from a rating of 3. The rating of 3 indicates 'neither agree nor disagree' on the five-point Likert scale adopted for this study.

Table 8-3: One-sample statistics for the results of taxonomy evaluation

	N	Mean	Std. Deviation	Sig. (2-tailed)
Perceived ease of use				
Learning to operate **Data Quality Assessment tool** is easy for me	10	4.1	0.56	<0.001
I find it easy to get **Data Quality Assessment tool** to do what I want	10	4.1	0.73	<0.001
It is easy to become skillful at using **Data Quality Assessment tool**	10	4.1	0.73	<0.001
Overall, I find **Data Quality Assessment tool** easy to use	10	4.4	0.51	<0.001
Satisfaction				
I am satisfied about the quality results I got after using the tool	10	4.5	0.52	<0.001
I am pleased for the overall quality of our data after using the tool	10	4.2	0.63	<0.001
I am content with the experience of using the tool	10	4.0	0.66	<0.001
How would you rate your overall satisfaction with us?	10	4.3	0.48	<0.001
Perceived usefulness				
Using this tool helps assess the quality of our system data.	10	4.0	0.66	<0.001
Using this tool increases my productivity in assessing and measuring the quality of our medical data.	10	4.2	0.78	<0.001
Using this tool enhances my effectiveness in managing and assessing the quality of our medical data.	10	4.2	0.63	<0.001
Overall, this tool is useful in assessing and assuring the quality of our medical data.	10	4.1	0.56	<0.001
Confirmation				
My experience with using this tool was better than I expected.	10	4.0	0.66	<0.001
The service level provided by this tool was better than what I expected.	10	4.1	0.56	<0.001

As shown in Table 8-3, the analysis results show that participants agreed on the practicality and usability of the proposed approach as the mean value of each DQ instrument evaluation construct was greater than the test values (3). The fact that all answers were significant, as p values for all DQ instrument evaluation constructs were less than 0.05, confirms that participants' perception were significantly positive towards the proposed DQ instrument.

Figure 8.14 implies that participants found the DQ instrument is easy to use and follow to conduct data quality activities. It also indicates that they perceived the usefulness of the proposed approach, and were satisfied with its results. Moreover, their expectation of the DQ instrument was met as the confirmation construct achieved 4.05 out of 5.

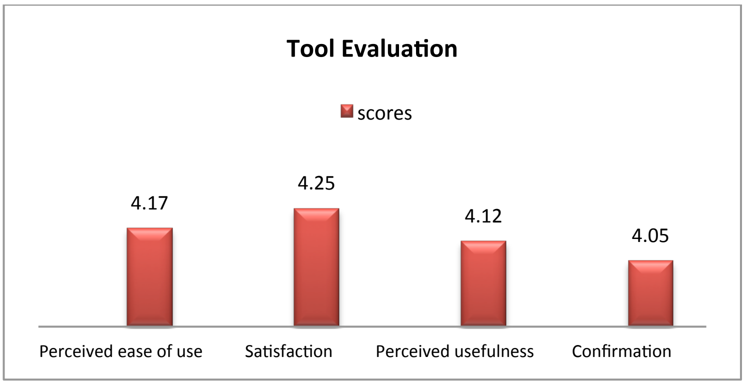


Figure 8.14: Mean of each scale of the tool evaluation

Figure 8.15 depicts the different perceptions of the information production roles towards the DQ instrument continuance intention. All roles (IT staff, health informatics and consultants) perceived almost the same level of usefulness of the new approach, achieving 4.06, 4.17 and 4.17 respectively. Interestingly, the consultants scored the construct 'perceived ease of use' higher than did the other roles, while IT staff scored it the lowest. This could be that the technical tasks of data quality assessment assigned to IT staff were time consuming and needed more effort than the tasks assigned to the other roles. With regard to satisfaction, all roles were satisfied with the results, but the health informatics, responsible for maintaining patient records, scored higher.

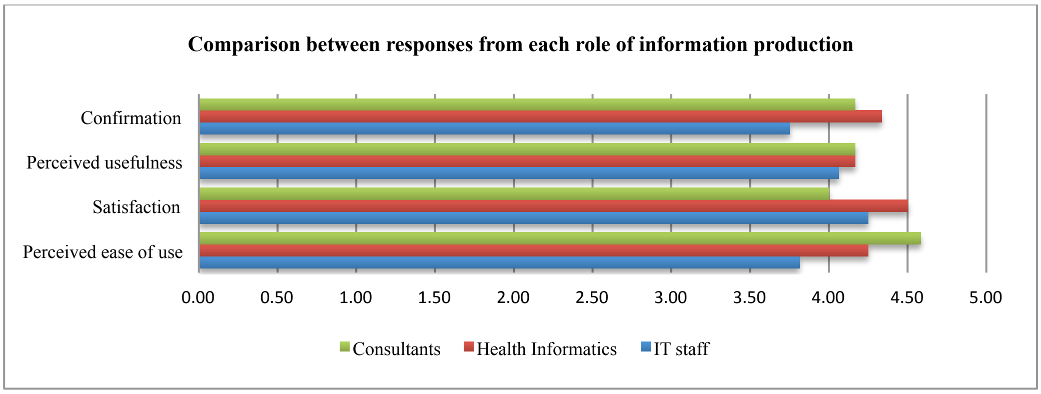


Figure 8.15: Mean comparison among roles in information production

8.3.1.3 Perceived ease of use

Figure 8.16 shows that the quality assessment team perceived that the DQ instrument was easy to use, as scores of all constructs were over 4. They showed no difficulty in operating the data quality assessment instrument in order to evaluate their own data.

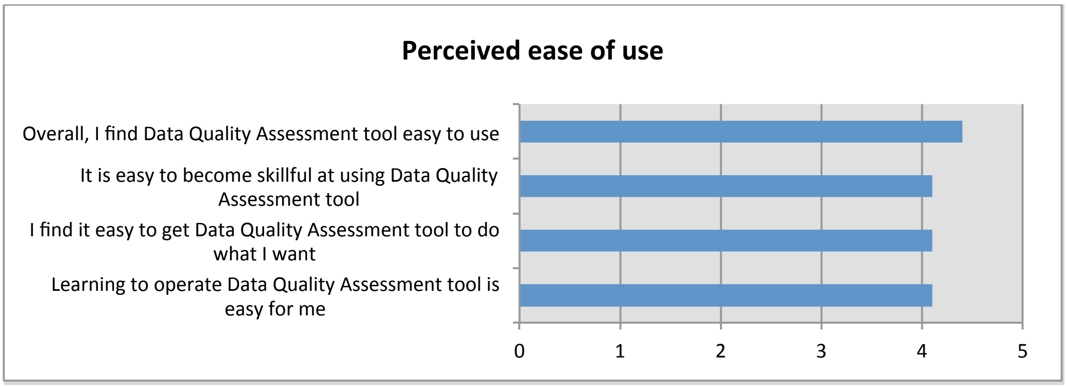


Figure 8.16: Output of the constructs of 'perceived ease of use'

8.3.1.4 Satisfaction

Using the DQ instrument to assess their data quality was satisfactory for the team, as shown in Figure 8.17. Importantly, the team was satisfied with the quality results of their data, achieving 4.5/5. This gives a good indication of the effectiveness of the DQ instrument. They were also content with the experience of using the DQ instrument. Overall, the quality assessment team showed a good level of satisfaction of the quality procedures before the assessment process, as well as the use of the DQ instrument.

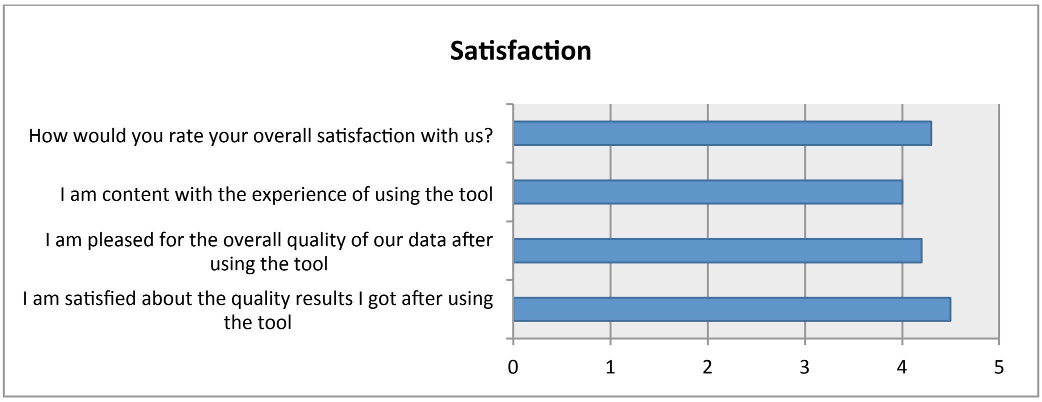


Figure 8.17: Output of the constructs of 'satisfaction'

8.3.1.5 Perceived usefulness

As shown in Figure 8.18, the items constructing 'perceived usefulness' indicated that the quality assessment team were convinced of the usefulness of the proposed DQ instrument. They believed that the DQ instrument would increase their productivity and enhance their effectiveness in assessing and managing the quality of their data. Overall, the team perceived the usefulness of the DQ instrument for assessing and assuring the quality if their data.

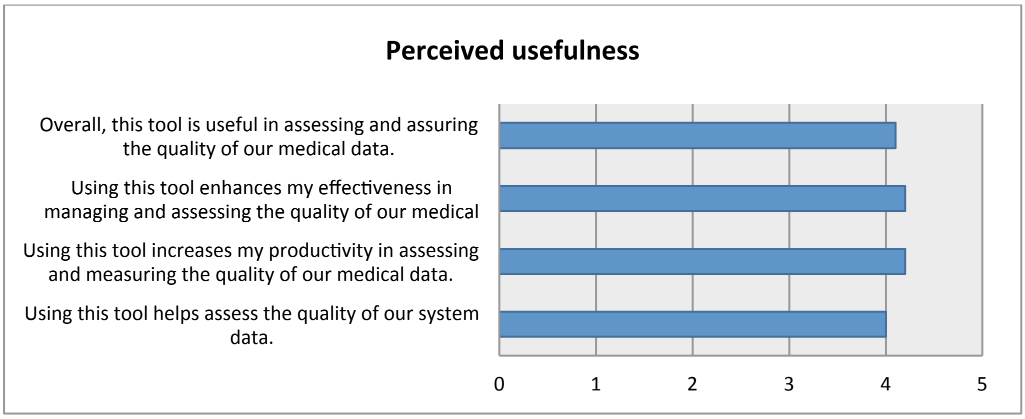


Figure 8.18: Output of the constructs of ’perceived usefulness’

8.3.1.6 Confirmation

This scale is users’ perception of congruence between expectation of the DQ instrument use and its actual performance. Figure 8.19 shows that their experience and the service level of the DQ instrument were better than expected, scoring 4.0 and 4.1 out of 5, respectively.

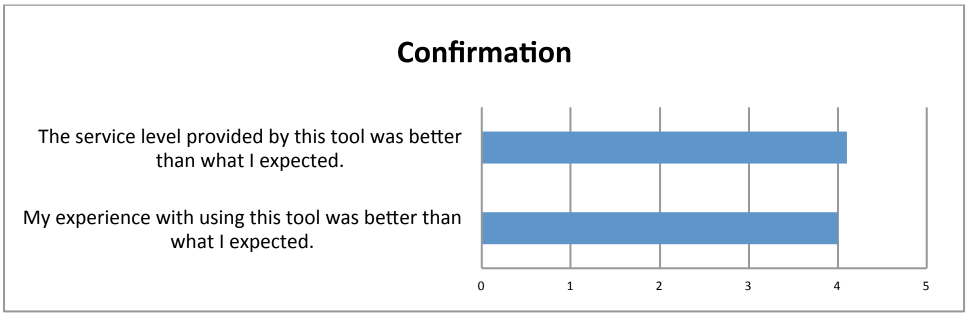


Figure 8.19: Output of the constructs of ’confirmation’

8.3.2 Managers’ views of the DQ instrument practicality

As described in Section 4.3.4, semi-structured interviews with two senior managers were conducted to discuss the results of the data quality assessment of their systems. The head of the IT department and the head of the health informatics department were exposed to the DQ instrument, as detailed in Section 8.2, along with the data quality issues raised by the quality assessment team. Some of the quality assessment team members attended the focus group-like semi-structured interview sessions.

Discussions focused on obtaining constructive feedback in relation to the practicality of the DQ instrument. Questions asked during the interviews covered topics such as:

- The current procedures of assuring the quality of their data

- Their perception of the quality of their data prior to the use of the DQ instrument
- Their perception of the quality level of their data after the results and findings of the case study
- The practicality and effectiveness of the DQ instrument.

With regard to the current procedures of data quality assurance, both seniors claimed that the existing procedures are acceptable at minimising the potential risk of poor data quality. These procedures include coding all diseases, procedures and treatments, and minimising free-text entries. However, there is no a systematic way to assess the data populating their system. They rely on individuals reporting any data quality issues to the IT department. The following quotes demonstrate their claims:

Head of IT: *We believe that data errors associated with data entries are the main contributor to poor data quality. We solved this issue by minimising free-text entries. We also applied not-null constraints to prevent poor data quality associated with incompleteness.*

Head of health informatics:

We worked with the IT department to assure the quality level of our data. We are working on a project to code all diseases and procedures. This would help the hospital staff not to make data errors by selecting from pre-defined entries.

They considered the quality level of their data as acceptable, since there were not many complaints received about data quality issues. However, they regarded data consumers as the largest contributor to data quality issues due to their role in data entering, editing and manipulating. This means that there is no systematic way to assess the quality of their data, as they rely on reports received from data consumers on data quality issues. Dimensions other than accuracy and completeness were not sufficiently present in their discussion of data quality issues during the first topic of the interviews.

After the results demonstration, the head of IT was satisfied with the findings associated with objectivity and security-related dimensions. However, the interviewee was not happy with the output of the measures of utility-related dimensions. The head of IT claimed that the system was designed and developed to meet the predefined requirements. The interviewee added that data consumers were the source of that low level of quality, as they are in charge of feeding the systems.

On the other hand, the head of heath informatics was happy with results, especially with the dimensions associated with objectivity. With regard to the findings of the utility-related dimensions, the head of heath informatics was expecting a lower level of quality, as he had already received many complaints from hospital staff about the system's usability. This raises the concern over communication between the system developers and the data consumers.

Both interviewees confirmed the usefulness of the DQ instrument. The DQ instrument promoted the awareness of some essential aspects of data quality. The head of IT emphasised that the measures used for assessing the quality level of our data covered some important issues, such errors associated with dual-language names.

Head of IT: *The tool is useful for capturing data quality problems, and made us aware of some important aspects of data quality such as relevance and interpretability. However, the process would be time consuming if the tool is manual.*

The head of heath informatics supported continuing to use this DQ instrument, since it involved important data quality dimensions that are usually overlooked by quality team. He believed that the DQ instrument is practical and effective.

8.4 Discussion of case study findings

The intent of the case study was to examine the practicality of the proposed data quality framework and its instrument in terms of their practicality and usefulness in assessing the data quality of EHR systems. The data quality framework consists of 11 dimensions that fit into three categories: objectivity, utility and security.

8.4.1 Data quality beyond accuracy and completeness

As discussed in Section 8.3.2, their initial perception of data quality was any quality issues with accuracy and completeness. This perception is common amongst health care professionals (Häyrinen et al., 2008). During the case study, EHR stakeholders were exposed to other crucial aspects of data quality. The individuals' awareness of data quality dimensions other than accuracy and completeness was increased through the data quality assessment process of their own sampled dataset. They managed to draw attentions to some quality issues on almost all dimensions including accuracy and completeness (see Section 8.2).

According to their feedback about the DQ instrument detailed in Section 8.3, they demonstrated good level of familiarity with the measures used to assess the quality of their dataset.

Moreover, the health provider of the facility where the case study was conducted has no systematic way of assessing the data quality of the data populating the EHR system, but works on an arbitrary basis. It relies on individual efforts rather than a systematic approach. The power of the systematic approach was demonstrated during the case study, as more data quality issues were captured. Some quality issues raised in Section 8.2 and 8.3.2 promote the need for better communication between developers and data consumers with regards to capturing requirements.

8.4.2 Helpfulness of the DQ instrument

The quality assessment team spent almost two months in assessing the quality of their real data. The output detailed in Section 8.2 shows that the DQ instrument managed to capture data quality problems that affect data quality. The DQ instrument helped the team to determine whether or not the data is of good quality. As confirmed in 8.3.2, the proposed data quality framework and its instrument assisted the team to cover more aspects of data quality in its assessment process. It also helped the team to pay more attention to the existence of dirty data.

The team members were asked to evaluate and rate the practicality of the DQ instrument through a questionnaire (see Section 4.3.4.2). They agreed on the effectiveness of the DQ instrument in assessing and managing the data quality as the mean of the representative question is 4.2 (see Figure 8.20 for all responses).

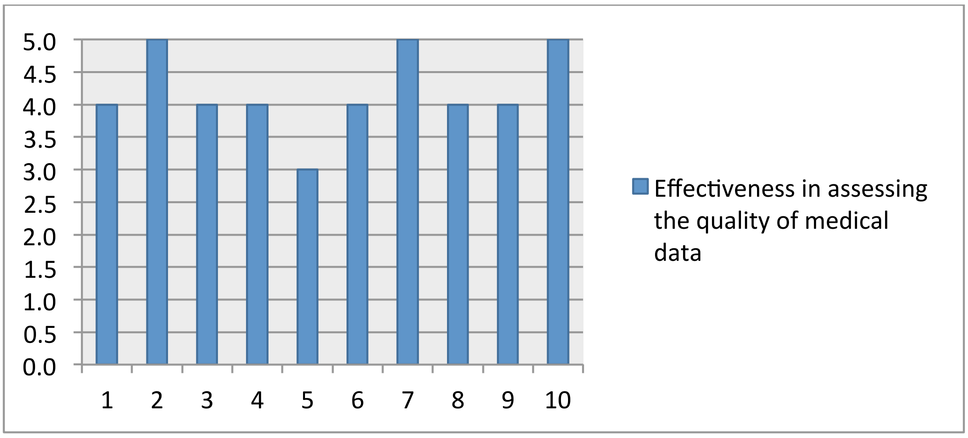


Figure 8.20: Quality team rating the practicality of the instrument

Moreover, the proposed framework and the DQ instrument encourage a systematic approach to data quality assessment. This would help an organisation to cover a wider range of data quality issues effectively. With the proposed framework, especially the objectivity-related dimensions and their associated measures, an organisation can partially apply some of the measures when considering only the specific needs of the organisation.

8.4.3 Practicality of the DQ instrument

In order to assess the practicality of the DQ instrument and the EHR stakeholders' intention to continue using the instrument, four determinants were introduced, namely perceived usefulness, perceived ease of use, satisfaction and confirmation. The quality assessment team involved in the case study evaluated the DQ instrument through the four determinants. The team confirmed the perceived usefulness of the DQ instrument. There was a statistically significant difference between the mean ratings of the representative question and the rating of 3 (see Table 8-3). Besides, two of the senior managers highlighted the potential of the DQ instrument for tackling data quality issues, and confirmed its perceived usefulness (see Section 8.3.2).

As shown in Table 8-3, the one sample t-test revealed a statistically significant difference between the mean rating of the question representing the determinant 'perceived ease of use' and the rating of 3. This confirmed the ease of using the DQ instrument, as perceived by the quality assessment team.

The data analysis in Section 8.3.1.4 shows that the team was satisfied with the instrument's performance. A rating of 4.5 for satisfaction is a good sign that the DQ instrument is satisfying in assessing the data quality. The interviews in Section 8.3.2 also revealed that the top management was well satisfied with the DQ instrument. Managers showed their contentment with the instrument's performance in highlighting the data quality issues in their systems and covering more data quality aspects.

As discussed in 8.3.1.6, the quality assessment team's expectations of the DQ instrument were confirmed and met. The service level provided by the DQ instrument was better than their expectations. The head of IT also confirmed that the service and the performance of the DQ instrument were beyond expectations, as it managed to draw attention to some issues that were overlooked by IT people.

8.4.4 Some issues with the DQ instrument

During the quality assessment process, the quality assessment team reported some observations of some significant quality issues which may compromise EHR system usability. They claimed that the DQ instrument did not manage to catch and discover these quality issues. However, all these issues could have been picked using the DQ instrument measures.

Table 8-4: Quality observations during the data quality assessment

Observation	Respective dimension	Respective system
- SOAP not available for inpatients	Usability – accessibility	SOAP
- History of drugs prescriptions is not available.	Usability – availability	Pharmacy
- Confusing drug names - Confusing drug doses	Interpretability	Pharmacy
- Test identification is only by date & time not by test name	Usability – retrievability	Lab
- Different abbreviations for one entity	Consistency	SOAP
- One tribe name had different spellings	Consistency	SOAP
- Some fields accidentally left without completion	Completeness – incomplete context	SOAP
- Volatility (discussed in Section 8.2.1)	Timeliness	LAB

Most of the observations were classified into the respective data quality dimensions and systems in Table 8-4. However, volatility was not captured through the DQ instrument. There is a good chance of improvement for the timeliness measures.

8.5 Chapter Summary

This chapter examines the usefulness and practicality of the proposed data quality framework and its associated DQ instrument. A real case study was conducted on a large hospital in Saudi Arabia. Twelve EHR stakeholders were involved in assessing the quality of dataset of 300 patients' records that were randomly retrieved.

Subsequently, the stakeholders were asked to evaluate their experience of the DQ instrument through a survey developed by the researcher. The survey measured four elements: 'the ease of use', 'the perceived usefulness', 'the user satisfaction' and 'the perception of congruence between expectation of the use and its actual performance'. After that, the outputs were discussed with two senior managers to examine their perception towards the proposed instrument.

The outcome of this case study is that the framework and its instrument show a good level of practicality, as the assessment results satisfied both the top management level and the EHR stakeholders representing all data production roles. The findings of the questionnaire survey confirm the practicality of the instrument.

Chapter 9 Conclusion and Future Work

This chapter highlights the aspects of our research on the data quality in relation to EHRs in Saudi Arabia. The main outcomes of the research are presented in Section 9.2. It next gives the directions for future research.

9.1 Conclusions

The aim of this research was to identify and develop a framework of data quality dimensions that can be used for assessing and improving the quality of data in Saudi EHR systems. This framework could be also used for prioritising improvement tasks that tackle issues associated with organizations' data. In addition, it focused on identifying a set of measures associated with each dimension s to facilitate the data quality assessment process for EHR systems.

In this research, the aspects of data quality proposed in the relevant literature were reviewed in order to develop a framework that represents characteristics and dimensions that impact upon data quality in EHR systems in Saudi Arabia. The identified framework consists of 11 clear data quality dimensions. These dimensions fall into three categories, which are **objectivity** (accuracy, completeness, consistency, timeliness), **utility** (usability, interpretability, relevance, provenance), and **security** (integrity, confidentiality and privacy). A questionnaire survey and semi-structured interviews confirmed the proposed data quality framework.

Subsequently, the data quality problems in data warehouse literature were studied in order to develop and produce initial measures for the objectivity dimensions. Initial measures for utility and security dimensions were also produced from the data quality literature. These measures were discussed and reviewed with two types of groups; data consumers and IT professionals. Due to some ambiguity with the measures associated with security dimensions, a further review was carried out with security experts to clear away the vagueness with some security measures.

After that, the practicality and usefulness of the DQ instrument, derived from the measures, were tested through a case study in a large hospital in Saudi Arabia. This included a quality assessment of 300 patients' records and an evaluation of the DQ instrument by 12 quality assessment team members.

9.2 Contribution to Knowledge

This thesis provided four contributions to knowledge. First, we developed a data quality framework applicable to EHR systems in Saudi Arabia. Second, the DQ instrument was developed to measure and assess the data quality in EHR settings in Saudi Arabia. Last, a new dimension-oriented taxonomy of data quality problems was introduced as a necessary stage in objectivity dimensions measures development process.

9.2.1 11-dimensional Data Quality Framework

We developed a data quality framework to address and answer the first research question RQ1:

'What data quality determinants are important for EHR stakeholders perceived data quality?'

An exploratory study was conducted with a group of five experts and another group of six data consumers, as well as confirmatory study with 66 EHR stakeholders. The result of these studies (Sections 5.1 and 5.2) confirms the 11 data quality dimensions. Therefore, the literature review, the result of the questionnaire and the findings of the interviews triangulate the conclusion that these 11 dimensions assess the quality of data populating EHR systems.

9.2.2 Dimension-Oriented Taxonomy of Data Quality Problems

As discussed in Section 3.2, a taxonomy of dimension-oriented data quality problems has been produced. Data quality problems were analysed and mapped into the most common data quality dimensions in the literature, which are accuracy, consistency, completeness and timeliness. Thus, the proposed taxonomy is concerned with identifying the problems from the perspective of quality dimensions, answering the sub-research question RQ2.1:

'What are the measuring items for objective data quality assessment?'

The proposed dimension-oriented data quality problems were discussed with experts and health professionals in order to confirm their relevance and to explore more quality problems that fall into the dimensions as yet not covered by this taxonomy (see Section 5.2.1.2).

The new proposed new taxonomy will help health organisations to prioritise the data quality problems associated with most desirable dimensions in the process of data quality assessment. Such a mechanism would facilitate the involvement of the data consumers at the assessment stage, as they are familiar with dimensions terminology, but not other, related works. The new taxonomy was examined and evaluated through a case study in a large hospital in Saudi Arabia.

9.2.3 Data quality instrument

The DQ instrument consists of 62 measures representing all dimensions identified in the framework discussed in Section 9.2.1 in order to be able to measure data quality in an EHR context.

- First, data quality items representing various facets of the 11 dimensions were identified to form an initial pool for the interviews. (RQ1 answered)
- Then, the 65 data quality measures were next discussed with two groups, experts and data consumers. The findings of the interviews (Sections 5.2.1.2 and 5.2.2.2) showed that most were approved as sound measures relevant to the respective dimensions. However, it was suggested that security-related items should be explored and discussed with experts in the field of information system security. In many cases, the researcher had to make decisions on the rationale, as there were discrepancies of opinion on some items. (RQ2.1 and RQ2.2 except security-related dimensions and their associated measures)
- Four security experts in the healthcare field reviewed and supplemented the security-related measures. (RQ2.2 fully answered)
- A case study was conducted in a large-scale hospital to examine the practicality of the instrument. (RQ2.3 answered)

9.2.4 Severity factors worsening data quality problems in EHR

It is noticeable that data quality problems become severe when associated with certain factors, meaning that more weight should be given to particular data quality items linked to specific factors. For example, a doctor would not accept an age-registered value of a patient aged 35 years as 60, but could tolerate a discrepancy of perhaps only three years. What makes quality problems severe is a high error rate. The severity factors identified with examples are shown in Table 9-1, addressing the research question RQ3:

‘What are the severity factors that make data quality problems more severe?’

Table 9-1: Severity factors with examples

Severity factor	Examples
Error distribution	20 errors found in 10 patients records is worse than if found in a patient record
Data sensitivity	Errors in clinical data is severer than if it is in demographic data
Data quality items	Misspelling error has less impact than erroneous error
Type of data collection tools	An error produced by a machine needs more attention than if it is entry mistake
Closeness to the correct values	If patient is aged 34, registered value of 60 is more severe than a registered value of 44
Cascading errors	An error leading to multi-errors is worse than isolated error

9.3 Future Research Directions

This Section suggests some interesting directions for future research. Some directions identified are briefly described as follows:

Research methodology – The confirmation of the data quality framework and its associated 62 measures were based on the literature review and an interpretation of mixed qualitative data, generated thorough a number of semi-structured interviews and one focus group discussions (see Section 5.2). Furthermore, there was some variance amongst the interviewees during the development of the proposed framework. These observations indicate that there may be other data quality dimensions (aspects) and their associated measures that need to be considered in the healthcare field. Hence, it is recommended to conduct further focus groups with health professionals and EHR stakeholders from different regions and settings in order to reveal some potentially additional data quality determinants that are important to data quality.

Due to the time constraints and the nature of the research scope, the DQ instrument was applied in only one large hospital. It is recommended to conduct more case studies to have deep insight on the practicality of the proposed instrument.

Data quality assessment (Weighting technique) – As displayed in Section 6.3, some data quality dimensions and associated measures vary in their importance, and data quality problems become severe when associated with certain factors. Therefore, the weighting technique may be useful to produce a precise quality score for each dimension; that is, more weight should be given to particular data quality dimensions and items linked to specific factors. This issue is known as multiple criteria decision making (MCDM), which refers to making decision in the presence of multiple attributes. Many potential methods have been proposed to solve MCDM problems. These include Analytic Hierarchy Process (AHP), Multiplicative Exponent Weighting (MEW) and Simple Additive Weighting (SAW) (Zanakis et al., 1998).

AHP is a powerful technique used for decision making, and allows the utility of quantitative and qualitative criteria in evaluation. It is a tool that breaks down a problem into sub-problems and then aggregates them into a conclusion. It facilitates organising the critical aspects of a problem into a hierarchically structured tree (Saaty, 1990). AHP utilises pairwise comparison to compare the importance of two dimensions on a subjective scale. Thus, this results in a matrix of importance according to the relative important given by the judges. In our study, two-stage AHP would be employed as pairwise comparison needs be applied on both dimension level and measure level.

Data quality assessment (dashboard-style) – According to the findings shown in Table F. 1 (APPENDIX F), there were two different sets of reliability level values; the data consumers provided two different sets of threshold values. As the threshold values are users' requirements, and specified by them, the mean of the two set was considered and could be amended by the actual users. The final threshold quality profile is presented in Table 9-2.

The defined measures of all dimensions can produce a dashboard-style data quality assessment scorecard. A measure could be assigned as Ignored, Low, Medium, High or Critical. Each of these categories has a different meaning and action of improvement.

Table 9-2: Data classification and reliability level of the dashboard-style quality score

Data set	Reliability level				
	Ignored	Low	Medium	High	Critical
Demographic data	<0.1%	<1%	<4%	<6%	>=6%
Clinical data	No error tolerated				
Administrative data	<0.1%	<1%	<4%	<6%	>=6%

Improvement stage – After assessment stage, the next step is to consider improvement steps for data quality issues that are identified in the assessment stage. Batini et al. (2009) have listed some well-known improvement strategies in the literature (see Table 9-3).

Table 9-3: Standard Strategies for Data Quality Improvement

Improvement strategy	Definition
Acquisition of new data	It replaces the bad quality data with higher-quality data
Standardization or normalization	It replaces nonstandard data values with values that comply with the standard.
Record linkage	It identifies and finds data set that refer to the same entity across different sources
Data and schema integration	It defines a unified view of data exist in heterogeneous data sources.
Source trustworthiness	It selects data sources based on their quality level.
Error localization and correction	It is mainly in statistical domain. It localises and removes data quality errors by detecting the records that do not meet a given set of quality rules.
Cost optimization	It defines cost-effective improvement strategies along with dimensions.
Process control	It introduces control checks and procedures during the data production process.
Process redesign	It is a long-term technique that redesigns processes in order to eliminate the root causes.

REFERENCES

Al-Shorbaji, N., 2001. *Health and medical informatics: technical paper*, World Health Organisation, Cairo.

Aldosari, B., 2014. Rates, levels, and determinants of electronic health record system adoption: A study of hospitals in Riyadh, Saudi Arabia. *International Journal of Medical Informatics*, 83(5), pp.330–342.

Almuayqil, S., Atkins, A.S. & Sharp, B., 2016. Ranking of E-Health Barriers Faced by Saudi Arabian Citizens, Healthcare Professionals and IT Specialists in Saudi Arabia. *Health*, 8(10), p.1004.

Aronson, J., 1994. A pragmatic view of thematic analysis. *Qualitative Report*, 2(1), pp.1–3.

Arts, D.G.T., De Keizer, N.F. & De Jonge, E., 2001. Data quality measurement and assurance in medical registries. In *Studies in Health Technology and Informatics*. Amsterdam: IOS Press, pp. 404–404.

Bah, S. et al., 2011. Annual survey on the level and extent of usage of electronic health records in government-related hospitals in Eastern Province, Saudi Arabia. *Perspectives in Health Information Management / AHIMA, American Health Information Management Association*, 8, p.1b. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3193507&tool=pmcentrez&rendertype=abstract.

Ballou, D.P. & Pazer, H., 1985. Modeling data and process quality in multi-input, multi-output information systems. *Management Science*, 31(2), pp.150–162.

Bates, D.W., 2000. Using information technology to reduce rates of medication errors in hospitals. *British Medical Journal*, 320(7237), pp.788–791.

Bates, D.W., Leape, L. & Cullen, D., 1998. Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. *JAMA: Journal of the American Medical Association*, 280(15), pp.1311–1316.

Batini, C. et al., 2009. Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, 41(3), p.16.

Batini, C. & Scannapieco, M., 2006. *Data quality: concepts, methodologies and* *techniques*, Springer.

Bayley, K.B. et al., 2013. Challenges in using electronic health record data for CER. *Medical Care*, 51, pp.S80–S86.

Beamon, B.M. & Ware, T.M., 1998. A process quality model for the analysis, improvement and control of supply chain systems. *Logistics Information Management*, 11(2), pp.105–113.

Begoyan, A., 2007. An overview of interopberability standards for electronic health records. In *Integrated Design and Process Technology, IDPT-2007.* Society for Design and Process Science, pp. 1–8.

Berg, B.L., 2011. *Qualitative Research Methods for the Social Sciences,* 8th edn. Harlow: Pearson.

Bhattacherjee, A., 2001. Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly*, 25(3), pp.351–370.

Blobel, B. & Pharow, P., 2009. Analysis and evaluation of EHR approaches. *Methods of Information in Medicine*, 48(2), p.162.

Bobrowski, M., Marré, M. & Yankelevich, D., 1999. A homogeneous framework to measure data quality. In *Proceedings of the International Conference on Information Quality*, Cambridge, MA, pp. 115–124.

Botsis, T. et al., 2010. Secondary use of EHR: Data quality issues and informatics opportunities. *AMIA Summits on Translational Science Proceedings AMIA Summit on Translational Science*, 2010, pp.1–5.

Bovee, M., Srivastava, R.P. & Mak, B., 2003. A conceptual framework and belief-function approach to assessing overall information quality. *International Journal of Intelligent Systems*, 18(1), pp.51–74.

Braun, V. & Clarke, V., 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), pp.77–101.

Brennan S., 2005. *The NHS IT Project:The biggest computer programme in the world... ever!.* Oxford: Radcliffe Publishing.

British Standards Institute, 2016. *BS EN ISO 27799: Health Informatics. Information security management in health using ISO/IEC 27002*, London: British Standards Institute.

British Standards Institute, 2011. *BS ISO 18308: Health informatics - Requirements for an electronic health record architecture.* London: British

Standards Institute.

Brown, S.M., 1997. Preparing Data for the Data Warehouse. In *Proceedings of the Conference on Information Quality*. Cambridge, MA, pp. 291–298.

Cabitza, F. & Batini, C., 2016. Information quality in healthcare. *Data and Information Quality*, pp.421–438.

Canada Health Infoway, 2009. *Building a Healthy Legacy Together – Annual Report 2008/2009*, Available at: www.infoway-inforoute.ca.

Canada Health Infoway, 2006. *Electronic Health Record Solution (EHRS) Blueprint*, Available at: https://www.infoway-inforoute.ca/en/component/edocman/391-ehrs-blueprint-v2-full/view-document.

Canadian Institute for Health Information, 2009. *The CIHI Data Quality Framework*, Available at: http://www.cihi.ca/CIHI-ext-portal/pdf/internet/DATA_QUALITY_FRAMEWORK_2009_EN.

Cappiello, C., Francalanci, C. & Pernici, B., 2003. Time-related factors of data quality in multichannel information systems. *Journal of Management Information Systems*, 20(3), pp.71–92.

Carey, M.A., 1994. The group effect in focus groups: Planning, implementing, and interpreting focus group research. In J. Morse (ed.), *Critical Issues in Qualitative Research Methods*. Thousand Oaks, CA: Sage, pp. 225–241.

Caro, A. et al., 2008. A proposal for a set of attributes relevant for Web portal data quality. *Software Quality Journal*, 16(4), pp.513–542.

Carson, C.S., 2000. What is data quality? A distillation of experience. In 9th Meeting of the Heads of National Statistical Offices of East Asian Countries, August, Japan. Available at: http://0374288.netsolhost.com/pdf/imf.pdf.

Cheung, C.S. et al., 2013. Factors associated with adoption of the electronic health record system among primary care physicians. *JMIR Medical Informatics*, 1(1), p.e1.

Churches, T., 2003. A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. *BMC Medical Research Methodology*, 3(1), p.1. Available at: http://www.biomedcentral.com/1471-2288/3/1.

Coiera, E., 2009. Building a national health IT system from the middle out. *Journal of the American Medical Informatics Association*, 16(3), pp.271–273.

Council of the European Union, 1995. *Article 6. European Union Data Protection Directive.* Available at: http://www2.warwick.ac.uk/fac/soc/law/elj/jilt/1996˙1/special/directive/#a2.1

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. P*sychometrika*, 16(3), pp.297–334.

Cykana, P., Paul, A. & Stern, M., 1996. DoD guidelines on data quality management. In *Proceedings of the 1996 Conference on Information Quality*. Cambridge, MA, pp. 154–171.

Dantu, R. et al., 2007. Securing medical networks. *Network Security*, 6, pp.13–16.

Davis, F.D., 1986. *A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results.* Massachusetts Institute of Technology.

Davis, F.D., Bagozzi, R.P. & Warshaw, P.R., 1989. User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), pp.982–1003.

Davis, N.A. & LaCour, M., 2014. *Health Information Technology*, 3rd edn., Elsevier Health Sciences.

DeLone, W.H. & McLean, E.R., 1992. Information systems success: The quest for the dependent variable. *Information Systems Research*, 3(1), pp.60–95.

Dolin, R.H. et al., 2006. HL7 clinical document architecture, release 2. *Journal of the American Medical Informatics Association*, 13(1), pp.30–39.

Dravis, F., 2004. Data quality strategy: A step-by-step approach. In *Proceedings of the Ninth International Conference on Information Quality (ICIQ-04).* Boston, MA: Institute of Technology, pp. 27–43.

Drever, E., 2003. Using semi-structured interviews in small-scale research: A teacher's guide, p.88. Available at: http://www.opengrey.eu/item/display/10068/423918 [Accessed 1 June 2014].

Eckerson, W., 2002. Data quality and the bottom line. *The Data Warehouse*

Institute, pp.1–32. Available at: http://download.101com.com/pub/tdwi/Files/DQReport.pdf.

Edwards Deming, W., 1982. *Out of the Crisis*. Cambridge, MA: Massachusetts Institute of Technology.

English, L.P., 1999. *Improving Data Warehouse and Business Information Quality: Methods for reducing costs and increasing profits*. New York: John Wiley.

Eppler, M.J. & Wittig, D., 2000. Conceptualizing information quality: A review of information quality frameworks from the last ten years. In *Proceedings of the 2000 Conference on Information Quality*, July, Cambridge, MA., pp. 83–96.

Esch, T. et al., 2016. Engaging patients through open notes: an evaluation using mixed methods. *BMJ Open*, 6(1), p.e010034.

Even, A. & Shankaranarayanan, G., 2007. Utility-driven assessment of data quality. *ACM SIGMIS Database*, 38(2), p.75.

Farzandipour, M. et al., 2010. Security requirements and solutions in electronic health records: Lessons learned from a comparative study. *Journal of Medical Systems*, 34(4), pp.629–642.

Fernández-Alemán, J.L. et al., 2013. Security and privacy in electronic health records: A systematic literature review. *Journal of Biomedical Informatics*, 46(3), pp.541–562.

Flick, U., 2002. *An Introduction to Qualitative Research*, Trent focus group, Nottingham. Available at: http://opac.rero.ch/get_bib_record.cgi?db=ne&#38;rero_id=R005035402.

Floridi, L., 2013. Information quality. *Philosophy & Technology*, 26(1), p.1.

Fox, C., Levitin, A. & Redman, T., 1994. The notion of data and its quality dimensions. *Information Processing & Management*, 30(1), pp.9–19.

Fragidis, L.L. & Chatzoglou, P.D., 2017. Development of Nationwide Electronic Health Record (NEHR): An international survey. *Health Policy and Technology*.

Gardyn, E., 1997. A data quality handbook for a data warehouse. In *Proceedings of the Conference on Information Quality*. Cambridge, MA, pp. 267–290.

Garets, D. & Davis, M., 2006. Electronic medical records vs. electronic health records: Yes, there is a difference. *Policy White Paper.* Chicago: HIMSS Analytics.

Ge, M. & Helfert, M., 2008. Data and information quality assessment in information manufacturing systems. *Business Information Systems.*

Gertz, M. et al., 2004. Report on the *dagstuhl* seminar. *ACM SIGMOD Record*, 33(1), pp.127–132.

Goodhue, D.L., 1995. Understanding user evaluations of information systems. *Management Science*, 41(12), pp.1827–1844.

Gray-Vickrey, P., 1993. Gerontological research use and application of focus groups. *Journal of Gerontological Nursing*, 19(5), pp.21–27.

Gray, D.E., 2009. *Doing research in the real world* 2nd edn., SAGE Publications.

Gritzalis, D. & Lambrinoudakis, C., 2004. A security architecture for interconnecting health information systems. *International Journal of Medical Informatics*, 73(3), pp.305–309.

Van Der Haak, M. et al., 2003. Data security and protection in cross-institutional electronic patient records. *International Journal of Medical Informatics*, 70(2), pp.117–130.

Haas, S. et al., 2011. Aspects of privacy for electronic health records. *International Journal of Medical Informatics*, 80(2), pp.e26--e31.

Hacker, K. et al., 2012. Impact of electronic health record transition on behavioral health screening in a large pediatric practice. *Psychiatric Services*, 63(3), pp.256–261.

Häyrinen, K., Saranto, K. & Nykänen, P., 2008. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *International Journal of Medical Informatics*, 77(5), pp.291–304.

Health Information and Quality Authority, 2011. *International Review of Data Quality.* Dublin: Health Information and Quality Authority. Available at: https://www.hiqa.ie/publications/international-review-data-quality.

Heard, S. et al., 2000. The benefits and difficulties of introducing a national approach to electronic health records in Australia. *Report to the Electronic Health Records Taskforce.* Adelaide: Commonwealth Department of Health and Aged Care.

Hennink, M., Hutter, I. & Bailey, A., 2010. *Qualitative Research Methods.*

London: Sage.

Hoffman, S. & Podgurski, A., 2008. Finding a cure: The case for regulation and oversight of electronic health record systems. *Harvard Journal of Law & Technology*, 22, p.103.

Hu, P.J. et al., 1999. Examining the technology acceptance model using physician acceptance of telemedicine technology. *Journal of Management Information Systems*, 16(2), pp.91–112.

Ilie, V. et al., 2009. Paper versus electronic medical records: The effects of access on physicians' decisions to use complex information technologies. *Decision Sciences*, 40(2), pp.213–241.

Institute of Medicine, 2000. *To Err is Human: Building a safer health system,* L.T. Kohn, J. M. Corrigan & M. S. Donaldson (eds), Institute of Medicine (US) Committee on Quality of Health Care in America. Washington, DC: National Academies Press. Available at: https://www.ncbi.nlm.nih.gov/pubmed/25077248.

Institute of Medicine (US) Committee on data standards for patient safety, 2003. *Key Capabilities of an Electronic Health Record System: Letter report*. Washington, DC: National Academies Press. Available at: http://www.nap.edu/openbook.php?record˙id=10781.

ISO, 2005. *20514:2005 Health Informatics-Electronic Health Record Definition, Scope and Context Standard.* Available at: http://www.iso.org/iso/iso˙catalogue/catalogue˙tc/catalogue˙detail.htm?csnumber=39525

Jahanbakhsh, M., Tavakoli, N. & Mokhtari, H., 2011. Challenges of EHR implementation and related guidelines in Isfahan. *Procedia Computer Science*, 3, pp.1199–1204.

Jain, A. et al., 2005. Responding to the rofecoxib withdrawal crisis: A new model for notifying patients at risk and their health care providers. *Annals of Internal Medicine*, 142(3), pp.182–186.

Jarke, M. et al., 1998. Architecture and quality in data warehouses. In B. Pernici & C. Thanos (eds), *Advanced Information Systems Engineering*, pp. 93–113. Springer-Verlag.

Jarke, M. et al., 2013. *Fundamentals of Data Warehouses.* Berlin: Springer Verlag.

Jarke, M. & Vassiliou, Y., 1997. Data warehouse quality: A review of the DWQ project. *Proceedings of the International Conference on Information Quality* (IQ), July, MIT, Cambridge, MA, pp. 299–313.

Jha, A.K. et al., 2008. The use of health information technology in seven nations. *International journal of medical informatics*, 77(12), pp.848–854.

Jick, T.D., 1979. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, 24(4), pp.602–611.

Jones, E.B. & Furukawa, M.F., 2014. Adoption and use of electronic health records among federally qualified health centers grew substantially during 2010–12. *Health Affairs*, 33(7), pp.1254–1261.

Juran, J.M., 1988. *Juran on Planning for Quality*. London: Collier Macmillan.

Kahn, M.G. et al., 2012. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Medical Care*, 50 Suppl(7), pp.S21-9.

Kalra, D., 2006. Electronic health record standards. *Yearbook of Medical Informatics,* pp.136-44.

Kerr, K., Norris, T. & Stockdale, R., 2007. Data quality information and decision making : A healthcare case study. In *18th Australasian Conference on Information Systems,* December, Toowoomba, Queensland, pp. 1017–1026.

Khalifa, M., 2013. Barriers to health information systems and electronic medical records implementation. A field study of Saudi Arabian hospitals. *Procedia Computer Science*, 21, pp.335–342.

Kim, G.R. & Lehmann, C.U., 2009. Electronic health records and interoperability for pediatric care. In C. U. Lehmann, G. R. Kim, & K. B. Johnson (eds), *Pediatric Informatics*. Health Informatics. New York: Springer, pp. 257–264.

Kim, W. et al., 2003. A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7(1), pp.81–99.

Kirkendall, E.S. et al., 2013. Transitioning from a computerized provider order entry and paper documentation system to an electronic health record: Expectations and experiences of hospital staff. *International Journal of Medical Informatics*, 82(11), pp.1037–1045.

Kitzinger, J., 1995. Qualitative research: Introducing focus groups. *British Medical Journal*, 311(7000), pp.299–302.

Klein, B.D. & Rossin, D.F., 1999. Data quality in neural network models: effect of error rate and magnitude of error on predictive accuracy. *Omega*, 27(5), pp.569–582.

Klein, R., 2007. An empirical examination of patient-physician portal acceptance. *European Journal of Information Systems*, 16(6), pp.751–760.

Kmietowicz, Z., 2004. Data collection is poor because staff don't see the point. *BMJ: British Medical Journal*, 328(7443), p.786.

Kovac, R., Lee, Y.W. & Pipino, L., 1997. Total Data Quality Management: The case of IRI. In *Proceedings of the International Conference on Information Quality* (IQ), July, MIT, Cambridge, MA, pp. 63–79.

Lee, Y.W. et al., 2002. AIMQ: A methodology for information quality assessment. *Information & Management*, 40(2), pp.133–146.

Lee, Y.W., 2003. Crafting rules: context-reflective data quality problem solving. *Journal of Management Information Systems*, 20(3), pp.93–119.

Lee, Y.W. & Strong, D.M., 2003. Knowing-why about data processes and data quality. *Journal of Management Information Systems*, 20(3), pp.13–39.

Leitheiser, R.L., 2001. Data quality in health care data warehouse environments. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*. pp. 1–10.

Levitin, A. & Redman, T., 1995. Quality dimensions of a conceptual view. *Information Processing & Management*, 31(1), pp.81–88.

Liaw, S.T. et al., 2012. Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. *International Journal of Medical Informatics*, pp.1–15.

van der Linden, H. et al., 2009. Inter-organizational future proof EHR systems: A review of the security and privacy related issues. *International Journal of Medical Informatics*, 78(3), pp.141–160.

Liu, L. & Ma, Q., 2005. The impact of service level on the acceptance of application service oriented medical records. *Information & Management*, 42(8), pp.1121–1135.

Longley, D. & Shain, M., 1987. Data and Computer Security: Dictionary of standards concepts and terms.

Lorence, D.P. & Jameson, R., 2002. Adoption of information quality management practices in US healthcare organizations: A national assessment. *International Journal of Quality & Reliabilty Management*, 19(6), pp.737–756.

Mandke, V. V & Nayar, M.K., 1997. Information integrity: A structure for its definition. In *Proceedings of the International Conference on Information Quality* (IQ), July, MIT, Cambridge, MA, pp. 314–338.

Matsumura, A. & Shouraboura, N., 1996. Competing with quality information. In *Proceedings of the International Conference on Information Quality* (IQ), July, MIT, Cambridge, MA, pp. 72–86.

McGuire, A.L. et al., 2008. Confidentiality, privacy, and security of genetic and genomic test information in electronic health records: points to consider. *Genetics in Medicine*, 10(7), pp.495–499.

McKinley, R.K. et al., 1997. Reliability and validity of a new measure of patient satisfaction with out of hours primary medical care in the United Kingdom: Development of a patient questionnaire. *British Medical Journal*, 314(7075), p.193.

McLafferty, I., 2004. Focus group interviews as a data collecting strategy. *Journal of Advanced Nursing*, 48(2), pp.187–194.

Meier, E., 2002. Medical privacy and its value for patients. In *Seminars in Oncology Nursing,* 18(2), pp. 105–108.

Merton, R.K., Fiske, M. & Kendall, P.L., 1990. *The Focused Interview: A manual of problems and procedures*, JSTOR. Available at: http://www.worldcat.org/oclc/464052735&ap=citavi.

Meyen, D. & Willshire, M.J., 1997. A data quality engineering framework. In *Proceedings of the Conference on Information Quality* (IQ), July, Cambridge, MA. pp. 95–116.

Moores, T.T., 2012. Towards an integrated model of (IT) acceptance in healthcare. *Decision Support Systems*, 53(3), pp.507–516.

Moreau, L. et al., 2008. The provenance of electronic data. *Communications of the ACM*, 51(4), pp.52–58.

Morgan, D.L., 1996. Focus groups. *Annual Review of Sociology*, pp.129–152.

Morrison, Z. et al., 2011. Understanding Contrasting Approaches to Nationwide Implementations of Electronic Health Record Systems: England, the USA and Australia. *Journal of Healthcare Engineering*, 2(1), pp.25–41.

Müller, H. & Freytag, J.-C., 2005. *Problems, Mmethods, and Challenges in Comprehensive Data Cleansing*. Humboldt-Univ. zu Berlin.

Murray, T.R., 2003. *Blending Qualitative and Quantitative Research Methods in Theses and Dissertations*. London: Sage.

National Alliance for Health Information Technology, 2008. *Report to the Office of the National Coordinator for Health Information Technology on Defining Key Health Information Technology Terms*, Available at: http://healthit.hhs.gov/portal/server.pt/gateway/PTARGS_0_10741_848133_0_0_18/10_2_hit_terms.pdf.

National E-Health Transition Authority, 2015. *My eHealth Record to National eHealth Record Transition Impact Evaluation*, Available at: http://www.digitalhealth.gov.au/get-started-with-digital-health/benefits/case-studies/northern-territory/918-myehr-to-national-ehealth-record-transition-impact-evaluation.

Naumann, F., 2002. Quality-driven query answering for integrated information systems, in Naumann (ed.), *Lecture Notes in Computer Science,* vol. 2261 Berlin: Springer Science & Business Media. ISBN: 978-3-540-43349-1 (Print) 978-3-540-45921-7 (Online)

Naumann, F. et al., 1999. Quality-driven integration of heterogeneous information systems. In *Proceedings of the International Conference on Very Large Data Bases*, September, Berlin, pp. 447–458.

Naumann, F. & Rolker, C., 2000. Assessment methods for information quality criteria. In *Proceedings of the 5th Annual Conference on Information Quality*, pp. 148–162.

Needham, D.M. et al., 2009. Improving data quality control in quality improvement projects. *International Journal for Quality in Health Care*, 21(2), pp.145–150.

Oliveira, P. & Rodrigues, F., 2005. A taxonomy of data quality problems. In *2nd International Workshop on Data and Information Quality* (in conjunction with CAISE'05), Porto, Portugal. pp. 219–233.

Orfanidis, L., Bamidis, P.D. & Eaglestone, B., 2004. Data quality issues in electronic health records: An adaptation framework for the Greek health system. *Health Informatics Journal*, 10(1), pp.23–36.

Orr, K., 1998. Data quality and systems theory. In *Communications of the ACM*. New York: ACM, pp. 66–71.

Osborn, R. & Schoen, C., 2013. The commonwealth fund 2013 international health policy survey in eleven countries. *Consulté le*, 19.

Paulson, L.D., 2000. Data quality: A rising e-business concern. *IT Professional*, 2(4), pp.10–14.

Pipino, L.L., Lee, Y.W. & Wang, R.Y., 2002. Data quality assessment. *Communications of the ACM*, 45(4), pp.211–218.

Polgar, S. & Thomas, S.A., 2011. *Introduction to Research in the Health Sciences,* 5th edn, Edinburgh: Churchill Livingstone.

Powell, R.A., Single, H.M. & Lloyd, K.R., 1996. Focus groups in mental health research: enhancing the validity of user and provider questionnaires. *International Journal of Social Psychiatry*, 42(3), pp.193–206.

Raghupathi, W. & Raghupathi, V., 2014. Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, 2(1), p.3.

Rahm, E. & Do, H.H., 2000. Data cleaning: Problems and current approaches. *IEEE Bulletin on Data Engineering*, 23(4), pp.3–13.

Redman, T.C., 1992. *Data Quality: Management and technology*, New York: Bantam.

Redman, T.C., 2001. *Data Quality: The Field Guide*, Digital Press.

Redman, T.C., 1996. *Data Quality for the Information Age*, Artech House.

Redman, T.C., 1998. The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2), pp.79–82.

Rigby, M. et al., 2001. Verifying quality and safety in health informatics services. *British Medical Journal*, 323(7312), pp.552–556.

Rindfleisch, T.C., 1997. Privacy, Information Technology, and Health Care. Communications of the ACM, 40(8), pp.92–100.

Rogers, Y., Sharp, H. & Preece, J., 2011. *Interaction Design: Beyond human-computer interaction,* 3rd edn., Chichester, West Sussex, UK: Wiley.

Rossman, G.B. & Rallis, S.F., 2003. *Learning in the Field: An introduction to* *qualitative research*. Thousand Oaks, CA: Sage.

Rubin, H.J. & Rubin, I.S., 2011. *Qualitative Interviewing: The art of hearing data.* Thousand Oaks, CA: Sage.

Saaty, T.L., 1990. How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, 48(1), pp.9–26.

Saunders, M., Lewis, P. & Thornhill, A., 2009. *Research Methods for Business Students*. Harlow: Prentice Hall.

Shortliffe, E.H. & Barnett, G.O., 2001. Medical data : Their acquisition , storage, and use. *Medical Informatics: Computer applications in health care and biomedicine*, 2nd edn, pp.45–82.

Starfield, B., 1994. Is primary care essential? *The Lancet*, 344(8930), pp.1129–1133.

Stracke, C. & Hildebrandt, B., 2007. Quality development and quality standards in e learning: Adoption, implementation, and adaptation. In *Proceedings of World Conference on Educational Multimedia Hypermedia and Telecommunication*. Association for the Advancement of Computing in Education (AACE), pp. 4158–4165.

Strong, D.M., Lee, Y.W. & Wang, R.Y., 1997. Data quality in context. *Communications of the ACM*, 40(5), pp.103–110.

Stufflebeam, D., 2001. Evaluation Models. *New Directions for Evaluation*, 2001(89), pp.7–98.

Su, Y. & Jin, Z., 2006. A methodology for information quality assessment in the designing and manufacturing process of mechanical products. *Information Quality Management: Theory and Applications*, pp.190–220.

Taggart, J. et al., 2012. The University of NSW electronic practice based research network: Disease registers, data quality and utility. In *Health Informatics: Building a Healthcare Future Through Trusted Information. Selected Papers from the 20th Australian National Health Informatics Conference* (HIC 2012). p. 219.

Tayi, G.K. & Ballou, D.P., 1998. Examining data quality. *Communications of the ACM*, 41(2), pp.54–57.

Thakkar, M. & Davis, D.C., 2006. Risks, barriers, and benefits of EHR systems: A comparative study based on size of hospital. *Perspectives in health information management / AHIMA, American Health Information*

Management Association, 3(5), pp.1–19.

Thong, J.Y.L., Hong, S.J. & Tam, K.Y., 2006. The effects of post-adoption beliefs on the expectation-confirmation model for information technology continuance. *International Journal of Human Computer Studies*, 64(9), pp.799–810.

Tupek, A., 2006. Definition of data quality. *Census Bureau Methodology & Standards Council, New York*, 6. Available at: http://repository.binus.ac.id/2009-1/content/M0094/M009468169.pdf [Accessed 15 June 2014].

Twinn, D.S., 1998. An analysis of the effectiveness of focus groups as a method of qualitative data collection with Chinese populations in nursing research. *Journal of Advanced Nursing*, 28(3), pp.654–661.

van der Veer, S.N. et al., 2010. Improving quality of care. A systematic review on how medical registries provide information feedback to health care providers. *International Journal of Medical Informatics*, 79(5), pp.305–323.

Venkatesh, V., 1999. Creation of favorable user perceptions: Exploring the role of intrinsic motivation. *MIS Quarterly*, 23(2), pp.239–260.

Wager, K.A., Glaser, J.P. & Lee, F.W., 2009. *Health Care Information Systems: A practical approach for health care management,* 2nd edn. London: John Wiley & Sons.

Wand, Y. & Wang, R.Y., 1996. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), pp.86–95.

Wang, R. & Strong, D., 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), pp.5–33.

Wang, R.Y., 1998. A product perspective on total data quality management. *Communications of the ACM*, 41(2), pp.58–65.

Wang, R.Y., Reddy, M.P. & Kon, H.B., 1995. Toward quality data: An attribute-based approach. *Decision Support Systems*, 13(3), pp.349–372.

Wang, R.Y., Storey, V.C. & Firth, C.P., 1995. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engineering*, 7(4), pp.623–640.

Weiskopf, N.G. et al., 2013. Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*, 46(5), pp.830–836.

Weiskopf, N.G. & Weng, C., 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1), pp.144–151.

Wimalasiri, J.S., Ray, P. & Wilson, C.S., 2004. Maintaining security in an ontology driven multi-agent system for electronic health records. In *Proceedings, 6th International Workshop on Enterprise Networking and Computing in Healthcare Industry – Healthcom 2004.* pp. 19–24.

Yi, M.Y. et al., 2006. Understanding information technology acceptance by individual professionals: Toward an integrative view. *Information & Management*, 43(3), pp.350–363.

Yin, R.K., 1992. The case study method as a tool for doing evaluation. *Current Sociology*, 40(1), pp.121–137.

Yoon-Flannery, K. et al., 2008. A qualitative analysis of an electronic health record (EHR) implementation in an academic ambulatory setting. *Informatics in Primary Care*, 16(4), pp.277–284.

Zanakis, S.H. et al., 1998. Multi-attribute decision making: A simulation comparison of select methods. *European Journal of Operational Research*, 107(3), pp.507–529.

Zhang, J. & Walji, M.F., 2011. TURF: toward a unified framework of EHR usability. *Journal of biomedical informatics*, 44(6), pp.1056–67.

Zheng, J. & Yu, H., 2016. Methods for linking EHR notes to education materials. *Information Retrieval Journal*, 19(1–2), pp.174–188.

Zhu, H. et al., 2014. *Data and information quality research: Its Evolution and Future*, London: Taylor and Francis Group LLC.

Zikmund, W.G. et al., 2013. *Business Research Methods.* Mason, OH: Cengage Learning.

Zmud, R.W., 1978. An Empirical investigation of the dimensionality of the concept of information. *Decision Sciences*, 9(2), pp.187–195.

APPENDIX A

UNIVERSITY OF Southampton

Questionnaire

Data quality dimensions in EHR

I am currently pursuing PhD research at the University of Southampton, United Kingdom. A key aim of my research is to explore and investigate the possible ways of assessing the quality of data in Electronic Health Record (EHR) systems, with a particular focus on Saudi health services. This survey is being conducted for PhD research on ”Data Quality in EHR” at the University of Southampton, United Kingdom. The survey aims to investigate the characteristics of data quality in the context of health information systems.

I would like your kind contribution in the research process by completing the questionnaire. I would also like to stress that anyone has the right to withdraw up until the final submission of their responses. All responses will be treated confidentially and respondents will be anonymous during the collection, storage and publication of research material. Responses are collected online and stored in a secure database.

Should you have any questions about the study or you wish to receive a copy of the results, please contact the researcher.

Omar Almutiry
Electronics and Computer Science (ECS)
The University of Southampton
Southampton
United Kingdom
Mail to: osa1a11@ecs.soton.ac.uk

Consent Form*

□ I have read and understood the information sheet (insert date /version no. of participant information sheet) and have had the opportunity to ask questions about the study. I also I agree to take part in this research project and agree for my data to be used for the purpose of this study. I understand my participation is voluntary and I may withdraw at any time without my legal rights being affected. I am happy to be contacted regarding other unspecified research projects. I therefore consent to the University retaining my personal details on a database, kept separately from the research data detailed above. The ’validity’ of my consent is conditional upon the University complying with the Data Protection Act and I understand that I can request my details be removed from this database at any time.

Data Protection *

□ I understand that information collected about me during my participation in this study will be stored on a password protected computer and that this information will only be used for the purpose of this study. All files containing any personal data will be made anonymous.

Demographic Information

1) **Gender**

● Male

○ Female

---

2) **What is the highest level of education you have completed?**

○ Diploma

○ Bachelor’s degree

○ Master’s degree

○ PhD

○ Other, Please specify:

---

3) **Do you work for health-related organisation?**

○ Yes

○ No

---

4) **Career**

○ Hospital doctor

○ GP

○ Nurse

○ Clinician

○ IT-related staff

○ Other, please specify:

---

5) **Where do you work?**

○ Hospital

○ Clinic

○ Health Ministry

○ Primary health centre

○ Other, please specify:

---

6) **How long have you been working in health-related field?**

○ 0-5 years

○ 6-10 years

○ More than 10 years

7) Does your organization store and process patient records electronically?

○ All

○ Some

○ None

○ Not applicable

Data Quality dimensions

In this Section, we are trying to identify and confirm the characteristics that affect the quality of data in Electronic Health Record (EHR) systems. The following questions would give the researcher an indication of how important to you these data quality characteristics are for health information systems:

8) Accuracy is defined as the extent to which registered data conforms to its actual value. How important for you is this characteristic?

○ Very important

○ Important

○ Neutral

○ Not important

○ Not at all important

9) Completeness is defined as the state in which information is not missing and is sufficient for the task. How important for you is this characteristic ?

○ Very important

○ Important

○ Neutral

○ Not important

○ Not at all important

10) Consistency is defined as that representation of data values remains the same in multiple data items in multiple locations. How important for you is this characteristic?

○ Very important

○ Important

○ Neutral

○ Not important

○ Not at all important

11) Relevance is defined as the extent to which information is appropriate and useful for the intended task. How important for you is this characteristic ?

○ Very important

○ Important

○ Neutral

○ Not important

○ Not at all important

12) Timeliness is defined as the state in which data is up to date and its availability is on time. How important for you is this characteristic ?

○ Very important

○ Important

○ Neutral

○ Not important

○ Not at all important

13) Usability is defined as the ease with which data can be accessed, used, updated, understood, maintained and managed. How important for you is this characteristic ?

○ Very important

○ Important

○ Neutral

○ Not important

○ Not at all important

14) Provenance is defined as the source of data, shown and linked to metadata about data. How important for you is this characteristic?

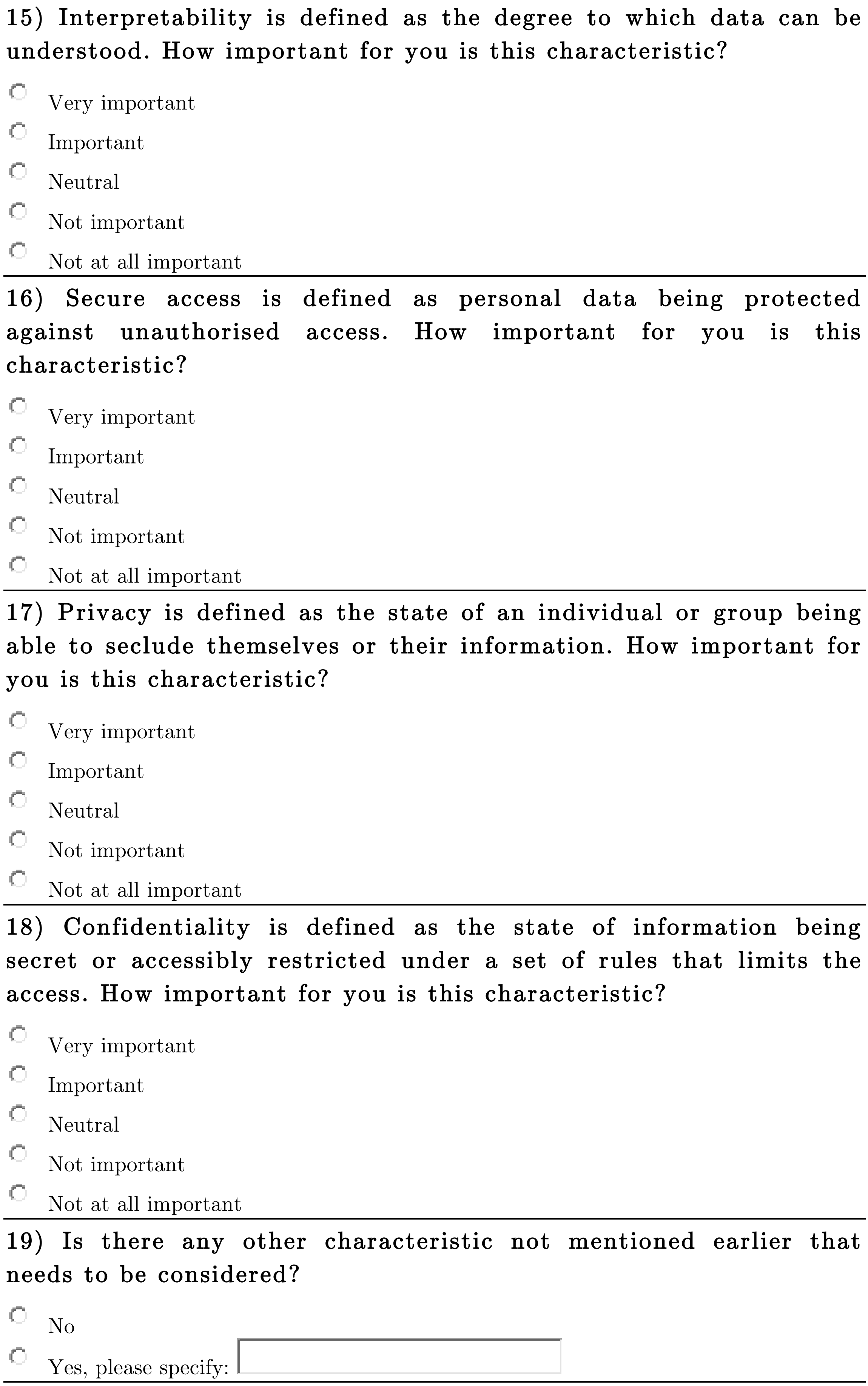
○ Very important

○ Important

○ Neutral

○ Not important

○ Not at all important



15) Interpretability is defined as the degree to which data can be understood. How important for you is this characteristic?

○ Very important

○ Important

○ Neutral

○ Not important

○ Not at all important

16) Secure access is defined as personal data being protected against unauthorised access. How important for you is this characteristic?

○ Very important

○ Important

○ Neutral

○ Not important

○ Not at all important

17) Privacy is defined as the state of an individual or group being able to seclude themselves or their information. How important for you is this characteristic?

○ Very important

○ Important

○ Neutral

○ Not important

○ Not at all important

18) Confidentiality is defined as the state of information being secret or accessibly restricted under a set of rules that limits the access. How important for you is this characteristic?

○ Very important

○ Important

○ Neutral

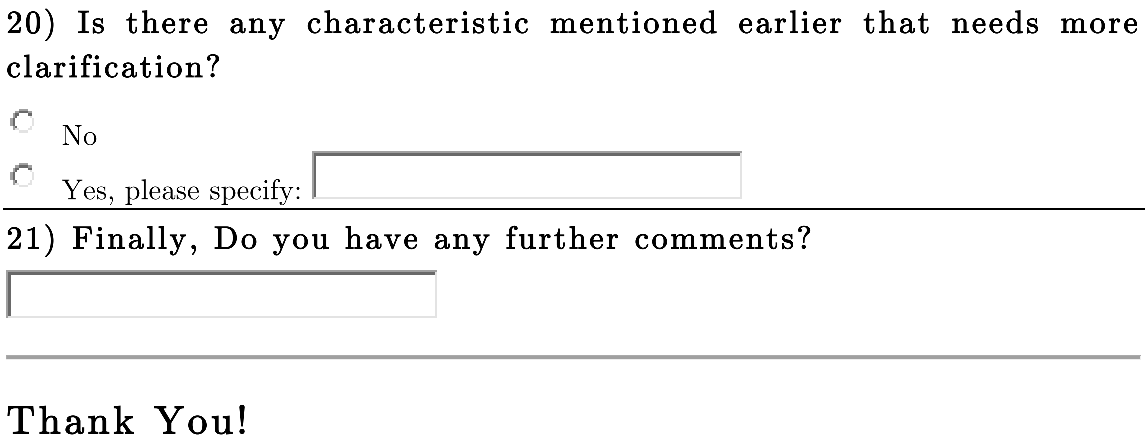
○ Not important

○ Not at all important

19) Is there any other characteristic not mentioned earlier that needs to be considered?

○ No

○ Yes, please specify: □



20) Is there any characteristic mentioned earlier that needs more clarification?

○ No

○ Yes, please specify:

21) Finally, Do you have any further comments?

Thank You!

APPENDIX B

APPENDIX C

Ethics reference number: ERGO/FPSE/9382	**Version:** 1	**Date:** 2014-03-05
Study Title: Data Quality Assessment		
Investigator: Omar Almutiry		

Please read this information carefully before deciding to take part in this research. If you are happy to participate you will be asked to sign a consent form. Your participation is completely voluntary.

What is the research about?

This is a student project, which aims to explore how to assure the quality of data. It addresses what and how to measure the quality of data populated in Health Information Systems. The ministry of Higher Education in Saudi Arabia funds the study. At the end of the study, I can email you my findings to see how your data was used, if you wish.

Why have I been chosen?

You have been approached because your department has sent requests to all employees for volunteers for my study, and it was optional.

What will happen to me if I take part?

You will first do sign a consent and then will be asked some questions with regard to my study. It will take about 40 mins in total.

Are there any benefits in my taking part?

The study will add to current knowledge about data quality in Health Information Systems.

Are there any risks involved?

There are no particular risks associated with your participation.

Will my participation be confidential?

All data collected is anonymous and your data will be kept confidential. It will be held on a password protected computer, and used only for the purposes of this study. In addition, the data will be anonymised by separating identifying data. It will be linked to your consent form by a code. It will be destroyed by the investigator once the study is done. If you would like to access your data after your participation, change it, or withdraw it, please contact the investigator (e-mail: osa1a11@ecs.soton.ac.uk) or the FoPSE Office (e-mail: lg11@soton.ac.uk) who will arrange this.

What happens if I change my mind?

You may withdraw at any time and for any reason. You may access, change, or withdraw your data at any time and for any reason prior to its destruction. You may keep any benefits you receive.

What happens if something goes wrong?

Should you have any concern or complaint, contact me if possible (investigator e-mail: osa1a11@ecs.soton.ac.uk), otherwise please contact the FoPSE Office (e-mail: lg11@soton.ac.uk) or any other authoritative body such as Dr Martina Prude, Head of Research Governance (02380 595058, mad4@soton.ac.uk).

Interview questions:

1- Does the definition provided reflect your understanding of "accuracy"?
2- How important to you is this dimension in measuring the quality of data?
3- Does the importance of this dimension vary along with the task at hand?
4- Does the importance of this dimension vary along with the class of the user?
5- Do you think each of these measuring items is considered a good measure for "accuracy"?
6- Do you think there is any overlapping between these items provided in the table below?
7- Are there any other measuring items that need to be considered?
8- From your point of view, do these measuring items have the same weight in assessing the information in terms of "accuracy"?

Accuracy: The extent to which registered data conforms to its actual value.	
Measurement items	
Schema Level	
Problems	**Example**
Illegal values due to invalid domain range	DoB= 30/02/1980
Instance Level	
Misspellings	Country= Girmany
Misfielded values	City= UK
Embedded values	Name= "J. Smith 12-01-2011 London"
Word transposition	Name1= "J. Smith" name2= "Muller K"
Wrong reference	Deptno= 12, found but incorrect
Erroneous entry	Age= 26, wrong real age= 36

Interview questions:

1- Does the definition provided reflect your understanding of "consistency"?
2- How important to you is this dimension in measuring the quality of data?
3- Does the importance of this dimension vary along with the task at hand?
4- Does the importance of this dimension vary along with the class of the user?
5- Do you think each of these measuring items is considered a good measure for "consistency"?
6- Do you think there is any overlapping between these items provided in the table below?
7- Are there any other measuring items that need to be considered?
8- From your point of view, do these measuring items have the same weight in assessing the information in terms of "consistency"?

<table>
<tr><td colspan="2"><b>Consistency</b>: Representation of data values remains the same in multiple data items in multiple locations.</td></tr>
<tr><th colspan="2">Measurement items</th></tr>
<tr><th colspan="2"><i>Schema Level</i></th></tr>
<tr><th>Problems</th><th>Example</th></tr>
<tr><td>Violated attribute dependencies</td><td>Age value not compatible with DoB value</td></tr>
<tr><td>Uniqueness violation</td><td>Uniqueness of student ID violated</td></tr>
<tr><td>Naming conflicts in multi-source</td><td></td></tr>
<tr><td>Structural conflicts in multi-source</td><td></td></tr>
<tr><td>Wrong categorical data</td><td>Not user-specified terms</td></tr>
<tr><td>Relational integrity violation</td><td>DeptNo=11, not found</td></tr>
<tr><th colspan="2"><i>Instance Level</i></th></tr>
<tr><td>Violated attribute dependencies</td><td>Postcode=SO171BJ city=London</td></tr>
<tr><td>Duplicated records in single/multi data source(s)</td><td>A student has 2 records</td></tr>
<tr><td>Contradicting records in single/multi source(s)</td><td>Same entity described differently in 2 records</td></tr>
<tr><td>Inconsistent spatial data</td><td></td></tr>
<tr><td>Different measure units in single/multi source(s)</td><td></td></tr>
<tr><td>Syntax inconsistency</td><td>2 dates with different format</td></tr>
</table>

Interview questions:

1- Does the definition provided reflect your understanding of “completeness”?
2- How important to you is this dimension in measuring the quality of data?
3- Does the importance of this dimension vary along with the task at hand?
4- Does the importance of this dimension vary along with the class of the user?
5- Do you think each of these measuring items is considered a good measure for “completeness”?
6- Do you think there is any overlapping between these items provided in the table below?
7- Are there any other measuring items that need to be considered?
8- From your point of view, do these measuring items have the same weight in assessing the information in terms of “completeness”?

<table>
<tr><td colspan="2"><b>Completeness</b>: The state in which information is not missing and is sufficient for the task.
Linkages between data promote the existence of further data.</td></tr>
<tr><th colspan="2">Measurement items</th></tr>
<tr><th colspan="2"><i>Schema Level</i></th></tr>
<tr><th>Problems</th><th>Example</th></tr>
<tr><td>Missing data where Null-not-allowed constraint enforced</td><td></td></tr>
<tr><th colspan="2"><i>Instance Level</i></th></tr>
<tr><td>Missing data where Null-not-allowed constraint not enforced</td><td>Data was unknown during initial stage</td></tr>
<tr><td>Missing record</td><td>A case not entered</td></tr>
<tr><td>Ambiguous data due to incomplete context</td><td></td></tr>
<tr><td>Semi-empty tuple</td><td></td></tr>
</table>

Interview questions:

1- Does the definition provided reflect your understanding of “timeliness”?
2- How important to you is this dimension in measuring the quality of data?
3- Does the importance of this dimension vary along with the task at hand?
4- Does the importance of this dimension vary along with the class of the user?
5- Do you think each of these measuring items is considered a good measure for “timeliness”?
6- Do you think there is any overlapping between these items provided in the table below?
7- Are there any other measuring items that need to be considered?
8- From your point of view, do these measuring items have the same weight in assessing the information in terms of “timeliness”?

Timeliness: The state in which data is up to date and its availability is on time.	
Measurement items	
Problems	**Example**
Outdated temporal value	Address not updated
Outdated reference	Foreign key not updated

Interview questions:

1- Does the definition provided reflect your understanding of "interpretability"?
2- How important to you is this dimension in measuring the quality of data?
3- Does the importance of this dimension vary along with the task at hand?
4- Does the importance of this dimension vary along with the class of the user?
5- Do you think each of these measuring items is considered a good measure for "interpretability"?
6- Do you think there is any overlapping between these items provided in the table below?
7- Are there any other measuring items that need to be considered?
8- From your point of view, do these measuring items have the same weight in assessing the information in terms of "interpretability"?

Interpretability: The degree to which data can be understood.	
Measurement items	
Problems	**Example**
Different representations due to use of abbreviation and cryptic values	
Different representations due to use of Alias/nickname	
Different representations due to use of encoding format	ASCII
Different representations due to use of special characters	Space, no space, dash

Interview questions:

1- Does the definition provided reflect your understanding of "interpretability"?
2- How important to you is this dimension in measuring the quality of data?
3- Does the importance of this dimension vary along with the task at hand?
4- Does the importance of this dimension vary along with the class of the user?
5- Do you think each of these measuring items is considered a good measure for "interpretability"?
6- Do you think there is any overlapping between these items provided in the table below?
7- Are there any other measuring items that need to be considered?
8- From your point of view, do these measuring items have the same weight in assessing the information in terms of "interpretability"?

USABILITY

Definition:

The ease with which data can be accessed, used, updated, understood, maintained and managed.

Measures:

1. Is this information easily accessible
2. Is this information easily retrievable
3. Is this information not usable
4. Is this information promptly accessible when needed
5. Is this information easily updatable
6. Is this information easily understood
7. Is this information usable
8. Is the availability of information for the patient treat adequate
9. Is this information easily manipulated

Interview questions:

1- Does the definition provided reflect your understanding of "relevance"?
2- How important to you is this dimension in measuring the quality of data?
3- Does the importance of this dimension vary along with the task at hand?
4- Does the importance of this dimension vary along with the class of the user?
5- Do you think each of these measuring items is considered a good measure for "relevance"?
6- Do you think there is any overlapping between these items provided in the table below?
7- Are there any other measuring items that need to be considered?
8- From your point of view, do these measuring items have the same weight in assessing the information in terms of "relevance"?

RELEVANCE
Definition:
The extent to which information is appropriate and useful for the intended task.
Measures:
1. Is this information relevant to the task at hand 2. Is this information useful to the task at hand 3. Is this information applicable to the task at hand 4. Is this information appropriate for the task at hand 5. Is this information irrelevant to the task at hand

Interview questions:

1- Does the definition provided reflect your understanding of "provenance"?

2- How important to you is this dimension in measuring the quality of data?

3- Does the importance of this dimension vary along with the task at hand?

4- Does the importance of this dimension vary along with the class of the user?

5- Do you think each of these measuring items is considered a good measure for "provenance"?

6- Do you think there is any overlapping between these items provided in the table below?

7- Are there any other measuring items that need to be considered?

8- From your point of view, do these measuring items have the same weight in assessing the information in terms of "provenance"?

PROVENANCE
Definition:
The source of data, shown and linked to metadata about data.
Measures:
1. Is the origin of this information clearly exist 2. Is this information owned by known subject 3. Is the creation date of this information shown 4. Is the update history of this information exist

Interview questions:

1- Does the definition provided reflect your understanding of “secure access”?
2- How important to you is this dimension in measuring the quality of data?
3- Does the importance of this dimension vary along with the task at hand?
4- Does the importance of this dimension vary along with the class of the user?
5- Do you think each of these measuring items is considered a good measure for “secure access”?
6- Do you think there is any overlapping between these items provided in the table below?
7- Are there any other measuring items that need to be considered?
8- From your point of view, do these measuring items have the same weight in assessing the information in terms of “secure access”?

SECURE ACCESS
Definition:
Personal data being protected against unauthorised access
Measures:
1. Is the access to this information sufficiently restricted 2. Does the access to this information require authentication process 3. Is the owner/creator of this information authorised 4. Is unauthorised access to this information sufficiently prevented

Interview questions:

1- Does the definition provided reflect your understanding of "confidentiality"?
2- How important to you is this dimension in measuring the quality of data?
3- Does the importance of this dimension vary along with the task at hand?
4- Does the importance of this dimension vary along with the class of the user?
5- Do you think each of these measuring items is considered a good measure for "confidentiality"?
6- Do you think there is any overlapping between these items provided in the table below?
7- Are there any other measuring items that need to be considered?
8- From your point of view, do these measuring items have the same weight in assessing the information in terms of "confidentiality"?

CONFIDENTIALITY
Definition:
The state of information being secret or accessibly restricted under a set of rules that limits the access
Measures:
1. Is the confidentiality of this information achieved 2. Is this information only accessible by authorised people 3. Is the access to this information granted only to persons who should see it 4. Is this information vulnerable 5. Is the sensitivity of this information clearly declared 6. Could this piece of information be disseminated without permission 7. In case of sharing, is there a clear consent for this information to be shared

Interview questions:

1- Does the definition provided reflect your understanding of "privacy"?
2- How important to you is this dimension in measuring the quality of data?
3- Does the importance of this dimension vary along with the task at hand?
4- Does the importance of this dimension vary along with the class of the user?
5- Do you think each of these measuring items is considered a good measure for "privacy"?
6- Do you think there is any overlapping between these items provided in the table below?
7- Are there any other measuring items that need to be considered?
8- From your point of view, do these measuring items have the same weight in assessing the information in terms of "privacy"?

PRIVACY

Definition:

The state of an individual or group being able to seclude themselves or their information.

Measures:

1. Is personally identifiable private information is being appropriately safeguarded
2. Is this personal identifiable information compliant with privacy policy
3. Is the policy regulation obeyed in this piece of information
4. Is privacy policy violated in this piece of information
5. Is the access to this personal identifiable information granted only to persons who should see it
6. In case of an individual being identified, is there a clear consent from this individual to be identified

General interview questions:

1- Do you have a methodology for assuring the quality of your data in your organisation?
2- Are there any other dimensions that need to be considered in our framework?
3- What are the factors that make the quality problems even severe, such as complexity and error distribution?
4- Any comments or suggestions

APPENDIX D

UNIVERSITY OF Southampton

Tool Evaluation

Department:

Code:

Date:

This questionnaire is designed to collect feedback about the use of the Data Quality Assessment tool. Your contribution is essential, and I will be grateful if you can complete this questionnaire and hand it back to the researcher.

Please rate each of the questions to correspond to experience of using the tool. All items are measured on a 1 to 5 scale where **1 is not at all** and **5 is completely**:

	Perceived ease of use					
1	Learning to operate **Data Quality Assessment** tool is easy for me.	1	2	3	4	5
2	I find it easy to get **Data Quality Assessment** tool to do what I want.	1	2	3	4	5
3	It is easy to become skillful at using **Data Quality Assessment** tool.	1	2	3	4	5
4	Overall, I find **Data Quality Assessment** tool easy to use.	1	2	3	4	5

	Satisfaction					
5	I am satisfied about the quality results I got after using the tool.	1	2	3	4	5
6	I am pleased for the overall quality of our data after using the tool.	1	2	3	4	5
7	I am content with the experience of using the tool.	1	2	3	4	5
8	How would you rate your overall satisfaction with us?	1	2	3	4	5

	Perceived usefulness					
9	Using this tool helps assess the quality of our system data.	1	2	3	4	5
10	Using this tool increases my productivity in assessing and measuring the quality of our medical data.	1	2	3	4	5

11	Using this tool enhances my effectiveness in managing and assessing the quality of our medical data.	1	2	3	4	5
12	Overall, this tool is useful in assessing and assuring the quality of our medical data.	1	2	3	4	5

	Confirmation					
11	My experience with using this tool was better than what I expected.	1	2	3	4	5
12	The service level provided by this tool was better than what I expected.	1	2	3	4	5

APPENDIX E

These are some templates used during the data quality assessment by the quality team members:

Usability assessment template

Data Quality Problem	Usability
Definition	The ease with which data can be accessed, used, updated, understood, maintained and managed.
Measures	
Item1	This information is easily accessible.
Item2	This information is easily retrievable.
Item3	This information is promptly accessible when needed.
Item4	This information is easily understood.
Item5	The availability of information for the patient treat is adequate.
Item6	This information is easy to manipulate to meet our needs.
Item7	In case of system failure, data is safely recoverable.

Please rate each of the metrics stated above for each medical data system. All items are measured on a 0 to 10 scale where **0 is not at all** and **10 is completely**:

	This information is easily accessible										
SOAP SYSTEM	0	1	2	3	4	5	6	7	8	9	10
LAB SYSTEM	0	1	2	3	4	5	6	7	8	9	10
PHARMACY SYSTEM	0	1	2	3	4	5	6	7	8	9	10

	This information is easily retrievable										
SOAP SYSTEM	0	1	2	3	4	5	6	7	8	9	10
LAB SYSTEM	0	1	2	3	4	5	6	7	8	9	10
PHARMACY SYSTEM	0	1	2	3	4	5	6	7	8	9	10

	This information is promptly accessible when needed										
SOAP SYSTEM	0	1	2	3	4	5	6	7	8	9	10
LAB SYSTEM	0	1	2	3	4	5	6	7	8	9	10
PHARMACY SYSTEM	0	1	2	3	4	5	6	7	8	9	10

	This information is easily understood

SOAP SYSTEM	0	1	2	3	4	5	6	7	8	9	10
LAB SYSTEM	0	1	2	3	4	5	6	7	8	9	10
PHARMACY SYSTEM	0	1	2	3	4	5	6	7	8	9	10

	The availability of information for the patient treat is adequate										
SOAP SYSTEM	0	1	2	3	4	5	6	7	8	9	10
LAB SYSTEM	0	1	2	3	4	5	6	7	8	9	10
PHARMACY SYSTEM	0	1	2	3	4	5	6	7	8	9	10

	This information is easy to manipulate to meet our needs										
SOAP SYSTEM	0	1	2	3	4	5	6	7	8	9	10
LAB SYSTEM	0	1	2	3	4	5	6	7	8	9	10
PHARMACY SYSTEM	0	1	2	3	4	5	6	7	8	9	10

	In case of system failure, data is safely recoverable										
SOAP SYSTEM	0	1	2	3	4	5	6	7	8	9	10
LAB SYSTEM	0	1	2	3	4	5	6	7	8	9	10
PHARMACY SYSTEM	0	1	2	3	4	5	6	7	8	9	10

Privacy assessment template

Data Quality Problem	Privacy
Definition	The right and desire of a person to control the disclosure of personal health information.
measures	
Item1	Patient's privacy is properly protected
Item2	Privacy policy is enforced for patient data access
Item3	In case of data release, anonymized data meets privacy protection
Item4	In case of an individual being identified, there is a clear consent from this individual to be identified

Please rate each of the metrics stated above for each medical data system. All items are measured on a 0 to 10 scale where **0 is not at all** and **10 is completely**:

	Patient's privacy is properly protected										
SOAP SYSTEM	0	1	2	3	4	5	6	7	8	9	10
LAB SYSTEM	0	1	2	3	4	5	6	7	8	9	10

PHARMACY SYSTEM	0	1	2	3	4	5	6	7	8	9	10

	Privacy policy is enforced for patient data access										
SOAP SYSTEM	0	1	2	3	4	5	6	7	8	9	10
LAB SYSTEM	0	1	2	3	4	5	6	7	8	9	10
PHARMACY SYSTEM	0	1	2	3	4	5	6	7	8	9	10

	In case of data release, anonymized data meets privacy protection										
SOAP SYSTEM	0	1	2	3	4	5	6	7	8	9	10
LAB SYSTEM	0	1	2	3	4	5	6	7	8	9	10
PHARMACY SYSTEM	0	1	2	3	4	5	6	7	8	9	10

	In case of an individual being identified, there is a clear consent from this individual to be identified										
SOAP SYSTEM	0	1	2	3	4	5	6	7	8	9	10
LAB SYSTEM	0	1	2	3	4	5	6	7	8	9	10
PHARMACY SYSTEM	0	1	2	3	4	5	6	7	8	9	10

Relevance assessment template

Data Quality Problem	Relevance
Definition	The extent to which information is appropriate and useful for the intended task.
Its Metrics	
Item1	This piece of medical information is useful to the task at hand.
Item2	This piece of medical information is applicable to the task at hand.
Item3	This piece of medical information is appropriate for the task at hand.

Please rate each of the metrics stated above for the medical data in hand. All items are measured on a 0 to 10 scale where **0 is not at all** and **10 is completely**:

	Record#1	Record#2	Record#3	Record#4	Record#5	Record#6	Record#7	Record#8	Record#9	Record#10	Record#11	Record#12	Record#13	Record#14	Record#15	Record#16	Record#17	Record#18	Record#19	Record#20
Item1	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10
Item2	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10
Item3	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10

	Record#21	Record#22	Record#23	Record#24	Record#25	Record#26	Record#27	Record#28	Record#29	Record#30	Record#31	Record#32	Record#33	Record#34	Record#35	Record#36	Record#37	Record#38	Record#39	Record#40
Item1	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10
Item2	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10
Item3	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10

	ord	ord	ord	ord	ord	ord	ord	ord	ord	ord	ord	ord	ord	ord	ord	ord	ord	ord	ord	ord

Item1	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10
Item2	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10
Item3	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10

ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.

Item1	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10
Item2	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10
Item3	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10

ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.	ord.#

Item1	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10
Item2	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10
Item3	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10	0 1 2 3 4 5 6 7 8 9 10

APPENDIX F

The data quality requirements analysis

The purpose of data quality requirements analysis is to survey stakeholders' opinions in order to identify quality issues and set new goals. This step is used to ascertain the critical areas that need improvement in EHR systems. It is to set an acceptability threshold at the dimensions level, in contrast to CIHI where the acceptance level is set at criteria level (Canadian Institute for Health Information, 2009).

In the interviews with experts and data consumers with regard to the data set of the measures, there was consensus that it should be categorised into three types: demographic, clinical and administrative. This would display the quality of scores for each type of data set. However, a DBA suggested adding 'financial data' as a fourth category. This proposal was not supported by the data consumers, especially the medical director, who claimed it is a part of administrative data.

With regard to the reliability levels, this issue was mainly discussed with the data consumers in order to analyse the quality assessment from the users' perspective, see Table F. 1. This introduced quality requirements where the organisation can specify acceptable quality values associated with each dimension. This quality profile would help the organisation to set their priorities in the improvement stage.

Table F. 1: Data consumers' feedback on multi-threshold values for the quality scores

Assurance level	Data set	Consult1	Consult2	Consult3	Med director	Consult4	Nurse
Ignored	Demographic	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%
	Clinical	No error tolerance	No error tolerance	No error tolerance	No error tolerance	No error tolerance	No error tolerance
	Administrative	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%
Low	Demographic	<1.0%	<1.0%	<1.0%	<1.0%	<1.0%	<1.0%
	Clinical	No error tolerance	No error tolerance	No error tolerance	No error tolerance	No error tolerance	No error tolerance
	Administrative	<1.0%	<1.0%	<1.0%	<1.0%	<1.0%	<1.0%
Medium	Demographic	<3.0%	<3.0%	<3.0%	<3.0%	<5.0%	<5.0%
	Clinical	No error tolerance	No error tolerance	No error tolerance	No error tolerance	No error tolerance	No error tolerance
	Administrative	<3.0%	<3.0%	<5.0%	<5.0%	<5.0%	<5.0%
High	Demographic	<5.0%	<5.0%	<5.0%	<5.0%	<7.0%	<7.0%
	Clinical	No error tolerance	No error tolerance	No error tolerance	No error tolerance	No error tolerance	No error tolerance
	administrative	<5.0%	<5.0%	<7.0%	<7.0%	<7.0%	<7.0%
Critical	Demographic	>5.0%	>5.0%	>5.0%	>5.0%	>7.0%	>7.0%
	Clinical	No error tolerance	No error tolerance	No error tolerance	No error tolerance	No error tolerance	No error tolerance
	Administrative	>5.0%	>5.0%	>7.0%	>7.0%	>7.0%	>7.0%