Toward A Framework For Data Quality In Cloud-Based Health Information System

Omar Almutiry, Gary Wills, Abdulelah Alwabel, Richard Crowder and Robert Walters
Electronics and Computer Science
University of Southampton
Southampton, UK
Email: {osa1a11,gbw,aa1a10,rmc,rjw1}@ecs.soton.ac.uk

Abstract—**This Cloud computing is a promising platform for health information systems in order to reduce costs and improve accessibility. Cloud computing represents a shift away from computing being purchased as a product to be a service delivered over the Internet to customers. Cloud computing paradigm is becoming one of the popular IT infrastructures for facilitating Electronic Health Record (EHR) integration and sharing. EHR is defined as a repository of patient data in digital form. This record is stored and exchanged securely and accessible by different levels of authorized users. Its key purpose is to support the continuity of care, and allow the exchange and integration of medical information for a patient. However, this would not be achieved without ensuring the quality of data populated in the healthcare clouds as the data quality can have a great impact on the overall effectiveness of any system. The assurance of the quality of data used in healthcare systems is a pressing need to help the continuity and quality of care. Identification of data quality dimensions in healthcare clouds is a challenging issue as data quality of cloud-based health information systems arise some issues such as the appropriateness of use, and provenance. Some research proposed frameworks of the data quality dimensions without taking into consideration the nature of cloud-based healthcare systems. In this paper, we proposed an initial framework that fits the data quality attributes. This framework reflects the main elements of the cloud-based healthcare systems and the functionality of EHR.**

Health Information System(HIS), Electronic Health Record (EHR), Data Quality (DQ), DQ Dimensions, Cloud Computing

I. INTRODUCTION

Electronic Health Record (EHR) refers to the digital form of a patient's medical record. It is defined as a repository of patient data in digital form. This record can be stored and exchanged securely and is accessible by different levels of authorized users (Häyrinen, Saranto, & Nykänen, 2008). Enhancing the quality of care is a noticeable advantage of adopting EHR systems. Many studies (Thakkar & Davis, 2006; Yoon-Flannery, 2008) have highlighted how such systems could enhance quality of care and support its continuity.

Cloud computing is a promising platform for health information systems in order to reduce costs and improve accessibility. Cloud computing represents a shift away from computing being purchased as a product to be a service delivered over the Internet to customers. Economic benefits are the key role behind the appearance of cloud computing (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009). The Cloud transforms IT assets from being capital expenditure to be operational expenditure. Traditionally, small and medium enterprises obtain IT infrastructure by purchasing it. In the cloud, using a server for five hours costs the same as using five servers for an hour (Armbrust et al., 2010).

Data quality in information systems and its 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 discussed. However, these frameworks have missed some important dimensions needed to ensure, for example, the integrity and origin of information (provenance). These missing dimensions are because the frameworks are generic and do not reflect the nature of the domain.

In the area of Health Information System, Data quality assurance is a challenging issue as the key barriers of optimally using data populated in cloud-based EHRs is the increasing data quantity with poor quality. "Fitness for use" is one of the best definitions of the data quality. This definition takes us even further beyond the traditional concerns with accuracy of data, as it will end up with many dimensions of data quality. So data quality is a concept with multi-dimensions.

Therefore, we developed an initial framework that concerns DQ in the context of cloud-based HIS. This framework is a result of filtering the existing data quality dimensions in many research, and checking their suitability to the nature of healthcare clouds.

This paper reviewed the notion of cloud computing, cloud-based EHR systems and their functionalities, and data quality. After that, it discussed the proposed framework and its life development. The paper concludes with discussion and future work.

II. HEALTHCARE CLOUD

In this section we briefly discus the notion of cloud computing and its potential in HIS. Then we briefly define the concept of personal Health Record (PHR), Electronic Medical Record (EMR) and Electronic Health Record (EHR). After

that, we study the functionalities of these systems which would help us identify the data quality dimensions used to measure and assess the quality of such systems.

A. Cloud computing and its attraction to healthcare IT

Cloud computing is a promising platform for EHR in order to reduce costs and improve accessibility. Cloud computing represents a shift away from computing being purchased as a product to be a service delivered over the Internet to customers. Economic benefits are the key role behind the appearance of cloud computing (Buyya et al., 2009). The Cloud transforms IT assets from being capital expenditure to be operational expenditure. Traditionally, small and medium enterprises obtain IT infrastructure by purchasing it. In the cloud, using a server for five hours costs the same as using five servers for an hour (Armbrust et al., 2010).

There are three common services delivered by Cloud: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). The underline infrastructure of a Cloud is consisted of one or more data centres, each has a massive number of computing resources. The IaaS delivery model allows users to acquire and release infrastructure resources (e.g. CPU and storage). PaaS offers a platform for developing end-to-end life cycle software development (Rimal, Choi, & Lumb, 2009) which contains development environment, set of applications to allow writing code, a set of ready packages to be used by other software and libraries (Hammond, Hawtin, Gillam, & Oppenheim, 2010). SaaS is a delivery model of applications provided by the Cloud to be run by Cloud users through web tools such as web services. This is the most abstract model of services, where users have no control over the Cloud infrastructure (Dillon, Wu, & Chang, 2010).

B. Cloud-based HIS-related challenges and issues

Some researches (Kuo, 2011; Zhang & Liu, 2010) highlighted some challenges and issues that could affect the adoption of this technology in healthcare field. The main concern is the lack trust in data security and privacy by users, the loss of governance and uncertain provider's compliance. This is due to the nature of this technology as it allows accessibility to different users. These issues will certainly affect the quality of data resided on cloud-based systems. The notion of cloud computing supports the accessibility from different sites and level of people. So there is a pressing need for assuring the quality of such system as it is a valuable source for the health stakeholders for their decisions.

C. The definitions of different types of healthcare systems

There are many terms that defined the patient-related electronic information in e-health services. These terms, EHR, EMR and EPR, are often used interchangeably in the healthcare filed despite the vital deference between these terminologies. Some people have confused EMR and EHR in spite of the fact that they describe the completely different concepts (Garets & Davis, 2006).

EMRs (Wager et al. 2009; Garets & Davis 2006) is a type of application environment composed 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 happened to patients during their encounters in the health care organisation. EMRs are widely used in North America and Japan but are regarded as outdated by many (Kim & Lehmann 2009).

Personal Health Record (PHR) is defined by some researchers (Alliance & Coordinator 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.

EHR refers to the digital form of a patient's medical record. It is defined as a repository of patient data in digital form. This record can be stored and exchanged securely and is accessible by different levels of authorized users (Häyrinen et al., 2008). What distinguishes EHR from EMR is that EHR combines electronic information of a patient from different care settings held in various healthcare systems.

D. The functionalities of healthcare systems

The Institute of Medicine (IOM) Committee in the USA (Hoffman & Podgurski, 2008) identified the key components of EHR systems and highlighted its functionalities. These core functionalities fall into eight categories, and are briefly discussed below:

- **Health Information and Data**: 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 by 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 like 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 health care. Communication can be among health care 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 health care system and among different systems (Thakkar & Davis, 2006).

- **Patient Support**: Many forms of patient support have shown significant effectiveness in health care in general. These forms include patient and family education and home telemonitoring.
- **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 health care units, but also in providing better service to patients.
- **Reporting and Population Health Management**: This feature makes the process of reporting less labour-intensive and time-consuming. It helps report patient safety and quality data and public health data.

III. DATA QUALITY

"Fitness for use" is one of the best definitions of the quality of data (Tayi & Ballou 1998), as this definition takes us beyond traditional concerns with data accuracy and with the many dimensions of data quality. Data quality includes not only data validation and verification, but also appropriateness of use (Orfanidis et al. 2004). Despite the fact that there are many frequently used dimensions such as accuracy, consistency, completeness, and timeliness, there is no consensus on a rigorously defined set of data quality dimensions (Strong et al. 1997; Tayi & Ballou 1998; Wand & Wang 1996).

A. Data Qualiy Dimensions

The definition of quality of data mentioned earlier states that data quality is a multi-dimensional concept. This definition implies that many other dimensions of data quality, including usefulness and usability, are important aspects of quality. Strong et al. (1997) classified these dimensions into four categories: intrinsic, accessibility, contextual and representational. Table 1 summarises some proposed data quality dimensions for information systems in general, along with their sources.

Table 1: Data quality dimensions in health information systems

Research	Data Quality Dimensions
(Ballou & Pazer, 1985)	Accuracy, completeness, consistency and timeliness.
(Strong et al., 1997)	Accuracy, objectivity, believability, reputation, accessibility, access security, relevancy, value-added, timeliness, completeness, amount of data, interpretability, ease of understanding, concise representation, consistent representation.
(Wang et al., 1995)	Accessibility, interpretability, usefulness, believability.
(Fox et al., 1994)	Accuracy, currentness, completeness, and consistency.
(Levitin & Redman, 1995)	Contents (relevance, unambiguous definitions, obtainability of values), scope (comprehensiveness, essentialness), level of details (attribute granularity domain precision), composition (naturalness, occurrence identifiability, homogeneity), consistency (semantic consistency, structural consistency) and reaction to change (robustness, flexibility).

B. Health-related Data Quality Dimensions

many researchers have defined data quality dimensions in the context of health. The Canadian Institute for Health Information (CIHI) defined five dimensions: accuracy, timeliness, comparability, usability and relevance. Each is divided into several characteristics, and each characteristic is divided further into criteria. Table 2 shows some frameworks of health-related data quality dimensions. Most common dimensions of data quality are accuracy, completeness, consistency, correctness and timeliness. However, Batini et al. (2009) claimed that the basic set of dimensions for data quality are accuracy, completeness, consistency and timeliness.

Table 2: Health-related Data Quality Dimensions

Research	Data Quality Dimensions
(Canadian Institute for Health Information, 2009)	Accuracy, timeliness, comparability, usability and relevance.
(Orfanidis et al., 2004)	Accessibility and availability, usability, security and confidentiality, provenance, data validation, integrity, accuracy and timeliness, completeness, and consistency.
(Liaw et al., 2012)	Accuracy, completeness, consistency, correctness and timeliness.

C. Impact of poor data quality

Enhancing the quality of care is a noticeable advantage of adopting EHR 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. Moreover, EHR promotes patient safety, as use of such systems improves patient safety by reducing medical errors in hospitals (Bates 2000; Bates et al. 1998). Medical errors can lead to death as, of which there are an estimated 98,000 each year in the United States, costing as much as $29 billion (Hoffman & Podgurski 2008). EHR systems could also notify patients about important changes in drug therapy (Jain et al. 2005).

IV. THE PROPOSED FRAMEWORK

The proposed framework was developed to tackle poor-quality data that compromise the quality of care. The proposed framework has three categories of health care data quality dimensions. These categories represent the main elements of e-health systems and healthcare systems. Development of this framework went through many stages to reflect the nature of cloud-based HIS.

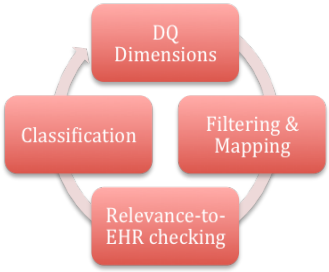

Fig. 1 The framework development process

Fig. 1 shows the process of developing the proposed framework. The process began with gathering data quality dimensions in organizations and health care systems. These dimensions were filtered to eliminate redundancies. In this step, literature review and dictionaries were used to avoid having two dimensions with the same implication. The next step was to check whether the dimension was relevant to EHR function, content and requirements. After that, the resulting dimensions were grouped into three categories: information, communication and security. These are considered the main elements of e-health care systems (Shoniregun et al. 2010).

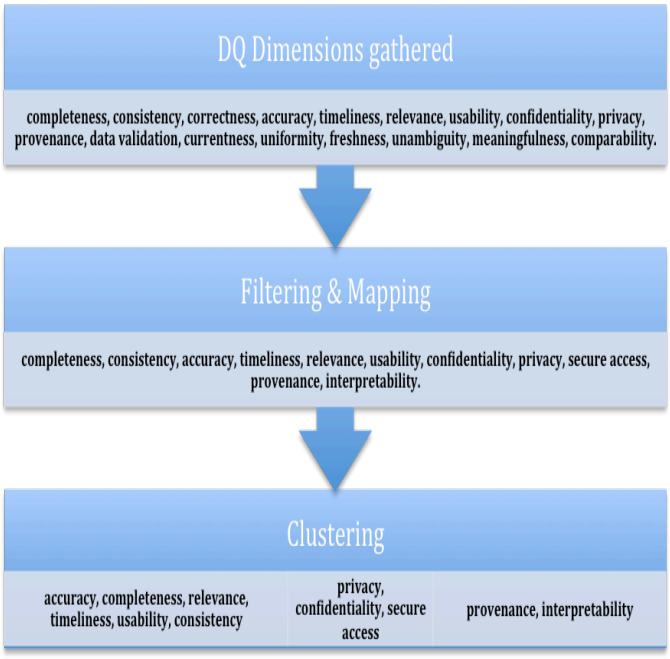

Fig. 2 The flow of the output during development

Fig. 2 shows the flow of reduction of the number of dimension at each stage. In the last stage, the dimensions are classified into three categories. This classification fits into our framework and covers all aspects of EHR systems, balancing comprehensiveness of dimensions with the nature of EHR systems. Fitting dimensions into the proposed framework gives a clearer definition of each dimension and helps identify what to measure and how.

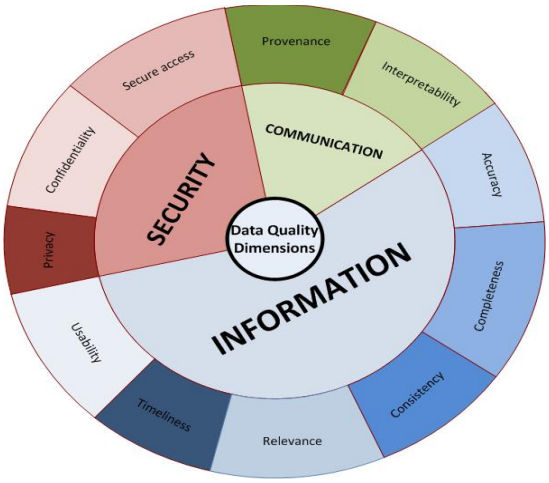

Fig. 3 The framework of data quality in cloud-based in health information systems

The characteristics of high-quality data fit into three categories: information, communication, and security. As can be seen from Fig. 3, there are 11 data quality dimensions in a framework of three categories. The following sections discuss the categories.

A. Information

Information is one of the three framework categories that shape e-health care systems. Most of existing approaches have addressed information-related dimensions. This category holds all dimensions associated with data characteristics, which are:

- **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.
- **Relevance**: The extent to which information is appropriate and useful for the intended task.
- **Timeliness**: The state in which data is up to date and its availability is on time.
- **Usability**: The ease with which data can be accessed, used, updated, understood, maintained and managed.

B. Communication

Communication is the second category of the framework. It concerns the correspondence between different care units. Because of this communication, EHR systems have multiple data items in multiple locations.

- **Provenance**: The source of data, shown and linked to metadata about data.
- **Interpretability**: The degree to which data can be understood.

C. Security

Security prevents personal data from being corrupted and controls access to ensure privacy and confidentiality.

V. CONCLUSION AND DISCUSSION

Cloud computing is a promising platform for health information systems in order to reduce costs and improve accessibility. However, adopting such technology arises the pressing need for assessing and measuring the quality of cloud-based health information systems. This is due to the fact that data quality is a multidimensional concept, and there is no consensus on rigorously defined set of data quality dimensions. This would emphasis the need of automating the mechanism of data quality measurement and semantic interoperability (Liaw et al., 2012).

Existing research focuses on data quality in generic information systems. These studies address data quality in many aspects aligned with data consumers. We developed an initial framework that concerns DQ in the context of electronic health care systems. This framework is a result of filtering the existing data quality dimensions in many research, and checking their suitability to the nature of e-health systems.

The next step will be examining and evaluating the proposed framework by conducting semi-structured interviews with EHR stakeholders in order to improve this work. Candidates for our research are IT professionals, GPs and health system managers in three general hospitals.

REFERENCES

[1] Alliance, T. N., & Coordinator, N. (2008). Defining Key Health Information Technology Terms. Health San Francisco, 299(03), 27–28. Retrieved from http://healthit.hhs.gov/portal/server.pt/gateway/PTARGS_0_10741_848133_0_0_18/10_2_hit_terms.pdf

[2] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., et al. (2010). A View of Cloud Computing. Communications of the ACM, 53(4), 50–58.

[3] Ballou, D. P., & Pazer, H. L. (1985). Modeling data and process quality in multi-input, multi-output information systems. Management science, 31(2), 150–162.

[4] Bates, D. W. (2000). Using information technology to reduce rates of medication errors in hospitals. BMJ, 320(7237), 788–791. doi:10.1136/bmj.320.7237.788

[5] Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems, 25(6), 599–616. doi:10.1016/j.future.2008.12.001

[6] Canadian Institute for Health Information. (2009). The CIHI Data Quality Framework.

[7] Dillon, T., Wu, C., & Chang, E. (2010). Cloud computing: Issues and challenges. 2010 24th IEEE International Conference on Advanced Information Networking and Applications (pp. 27–33). Ieee. doi:10.1109/AINA.2010.187

[8] DW, B., LL, L., DJ, C., & et al. (1998). EFfect of computerized physician order entry and a team intervention on prevention of serious medication errors. JAMA: The Journal of the American Medical Association, 280(15), 1311–1316. doi:10.1001/jama.280.15.1311

[9] Fox, C., Levitin, A., & Redman, T. (1994). THE NOTION OF DATA AND ITS quality dimensions. Information Processing & Management, 30(I), 9–19.

[10] Garets, D., & Davis, M. (2006). Electronic Medical Records vs . Electronic Health Records : Yes , There Is a Difference. HIMSS Analytics, 1–14. Retrieved from http://www.stttelkom.ac.id/staf/apk/riset/2012/EHMS/pustaka/ehealth/emr vs ehr.pdf

[11] Hammond, M., Hawtin, R., Gillam, L., & Oppenheim, C. (2010). Cloud computing for research. Final Report. Curtis+ Cartwright Consulting Ltd, 7.

[12] Hoffman, S., & Podgurski, A. (2008). Finding a Cure: The Case for Regulation and Oversight of Electronic Health Record Systems. Harv. JL & Tech., 22, 103.

[13] 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), 291–304.

[14] Jain, A., Atreja, A., & Harris, C. (2005). Responding to the Rofecoxib Withdrawal Crisis : A New Model for Notifying Patients at Risk and Their Health Care Providers. Annals of internal …, 182–187.

[15] 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 (pp. 257–264). Springer New York. Retrieved from http://dx.doi.org/10.1007/978-0-387-76446-7_18

[16] Kuo, A. (2011). Opportunities and Challenges of Cloud Computing to Improve Health Care Services. Journal of Medical Internet Research. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/pmc3222190/

[17] Levitin, A., & Redman, T. (1995). Quality dimensions of a conceptual view. Information Processing & Management, 31(1), 81–88. doi:10.1016/0306-4573(95)80008-H

[18] Liaw, S. T., Rahimi, A., Ray, P., Taggart, J., Dennis, S., De Lusignan, S., Jalaludin, B., 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, 1–15. doi:10.1016/j.ijmedinf.2012.10.001

[19] Orfanidis, L., Bamidis, P., & Eaglestone, B. (2004). Data Quality Issues in Electronic Health Records: An Adaptation Framework for the Greek Health System. Health Informatics Journal.

[20] Parker, M., Stofberg, C., Harpe, R. D. la, Venter, I., & Wills, G. (2006). Data quality: How the flow of data influences data quality in a small to medium medical practice.

[21] Rimal, B. P., Choi, E., & Lumb, I. (2009). A Taxonomy and Survey of Cloud Computing Systems. 2009 Fifth International Joint Conference on INC, IMS and IDC, 44–51. doi:10.1109/NCM.2009.218

[22] Shoniregun, C. A., Dube, K., & Mtenzi, F. (2010). Electronic healthcare information security. Springer.

[23] Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. Communications of the ACM, 40(5), 103–110.

[24] Tayi, G., & Ballou, D. (1998). Examining Data quality. Communications of the ACM.

[25] 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.

[26] Wager, K. A., Lee, F. W., & Glaser, J. P. (2009). Health Care Information Systems: A Practical Approach for Health Care Management (p. 5). Wiley. Retrieved from http://books.google.co.uk/books?id=n0X0dTc4o_kC

[27] Wand, Y., & Wang, R. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM.

[28] Wang, R. Y., Reddy, M. P., & Kon, H. B. (1995). Toward quality data: An attribute-based approach. Decision Support Systems, 13(3), 349–372.

[29] Yoon-Flannery, K. (2008). A qualitative analysis of an electronic health record (EHR) implementation in an academic ambulatory setting. … in primary care, 277–284.

[30] Zhang, R., & Liu, L. (2010). Security models and requirements for healthcare application clouds. Cloud Computing (CLOUD), 2010 IEEE 3rd ….