IADIS International Journal on WWW/Internet
Vol. 13, No. 2, pp. 98-114
ISSN: 1645-7641

DIMENSION-ORIENTED TAXONOMY OF DATA QUALITY PROBLEMS IN ELECTRONIC HEALTH RECORD

Omar Almutiry, Gary Wills and Richard Crowder. *Department of Electronics and Computer Science, University of Southampton,Southampton, UK.*

ABSTRACT

The provision of high quality data is of considerable importance to health sector. Healthcare is a domain in which the timely provision of accurate, current and complete patient data is one of most important objectives. The quality of Electronic Health Record (EHR) data concerns health professionals and researchers for secondary use. To ensure high quality data in health sector, health-related organisations need to have appropriate methodologies and measurement processes to assess and analyse the quality of their data. Yet, no adequate attention has been paid to the existing data quality problems (dirty data) in health-related research. In practice, anomalies detection and cleansing is time-consuming and labour-intensive which makes it unrealistic to most health-related organisations. This paper proposes a dimension-oriented taxonomy of data quality problems. The mechanism of the data quality assessment relates the business impacts into data quality dimensions. As a case study, the new taxonomy-based data quality assessment was used to assess the quality of data populating an EHR system in a large Saudi Arabian hospital. The assessment results were discussed and reviewed with the top management of the hospital as well as the assessment team who participated in the data quality assessment process. Then, the assessment team evaluated this new approach.

KEYWORDS

Data Quality, Information Quality, Quality Problem, Dirty Data, Data Quality Dimensions Electronic Health Record (EHR).

1. INTRODUCTION

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 recognized 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).

Electronic Health Records (EHRs) are a digital form of patient medical records that surpass many existing registries and repositories. 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 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. The definition used in this study for the quality of data is its 'fitness for use'. This definition brings up a concern beyond traditional concerns with data accuracy: that it will lead to many dimensions of data quality, making data quality a multi-dimensional concept.

The healthcare field is known as information-intensive, since massive data information is generated on a daily basis. An estimated 30 per cent of the health budget usually goes to issues related to information handling (Health Information and Quality Authority, 2011). 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.

In this paper, a new classification of dimension-oriented data quality problems is proposed and discussed with EHR stakeholders and IT. This devotes the importance of end-users involvement to capture the their requirements and needs. The new taxonomy is to be validated as a feasibility test for the proposed measurement approach.

The remainder of this paper is as follows. In section 2, a review of data quality issues is presented. Section 3 presents the taxonomy of dimension-oriented data quality problems. Confirmation of the new taxonomy in the context of EHR is presented in section 4. Section 5 highlights the applicability and practicality side of the new taxonomy by conducting a case study on a large health provider. Finally, the paper is concluded in section 6.

2. DATA QUALITY ISSUES

The importance of high quality health information for healthcare decisions is widely recognised by many health-related bodies through their initiatives (Batini et al., 2009). The Canadian Institute for Health Information (CIHI) defined data quality in the context of users; that is, if data satisfies users' needs, then it is fit for use (Canadian Institute for Health Information, 2009).

2.1 Data Quality Dimensions

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 between existing dimensions (Weiskopf & Weng, 2013). However, (Weiskopf & Weng 2013) largely reviewed clinical research literature for the data

quality dimensions in the context of EHR, and identified five dimensions, which are completeness, correctness, concordance, plausibility and currency. These findings go along with widely common dimensions in the literature of data quality. These are accuracy, consistency, completeness and timeliness (Batini et al. 2009; Liaw et al. 2012). These dimensions represent the basic set for data quality, and broadly accepted.

In literature, there are different classifications of quality dimensions with a number of discrepancies in the definitions of most (Batini et al., 2009). 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 emphasize the importance of providing a design-oriented definition of data quality that will reflect the nature of information systems.

Thus, this paper adopted the basic set of dimensions due to it being widely accepted. Besides, 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 assessment.

The definitions of these dimensions were discussed with IT experts and health professionals. It concludes with the following definitions (Almutiry et al. 2013):
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.

2.2 Data Quality Problems and Dirty Data

Organisations and enterprises tend not to pay 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.

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.

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.

Kim et al. (2003) take another pattern of classification of 'dirty data'. They look at dirty data 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.

This is a recent work compared to those mentioned above. Researchers Oliveira and Rodrigues (2005) present a comprehensive taxonomy of data quality problems through reviewing the previous work (Kim et al. 2003; Rahm & Do 2000; Müller & Freytag 2005), using a bottom-up approach, from the lowest level where data quality can appear to the highest level where these problems also exist. This approach resulted in six levels of granularity ranging from problems exist in single attribute value (lowest level) to problems exist in multi-source (highest level).

2.3 Data Quality and Dirty Data Taxonomies

In the literature, data quality problems and 'dirty data' were used in interchangeably, addressing issues and problems that lead to poor quality of data and, consequently, produce unreliable data. (Müller & 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.

Rahm & Do (2000), Kim et al. (2003) and Oliveira & Rodrigues (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 in their research a comprehensive hierarchal taxonomy that captures 33 types of 'dirty data' of 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'.

3. DIMENSIONS-ORIENTED TAXONOMY OF DATA QUALITY PROBLEMS

As discussed earlier, data quality is multi-dimensional concept. It is widely defined as 'fitness for use', emphasising the importance of data consumer's perspective of quality. Researchers since the early 1990s,have identified many dimensions that capture different facets of data quality. Some of these works were empirically proven (Wang & Strong 1996). Rationally, data quality problems should fall under the dimensions proposed in the literature. However, there is no taxonomy of 'dirty data' that groups quality problems based on their relation to dimension measurement. In this section, a dimension-oriented taxonomy of data quality problems is proposed, and that would help organisations assess each dimension of quality and prioritise them.

Figure 1 illustrates the taxonomy proposed by Rahm and Do (2000) was adopted as an initial collection of data quality problems. Subsequently, other types of 'dirty data' found in other studies were thoroughly analysed and filtered to identify those not included in that initial collection.

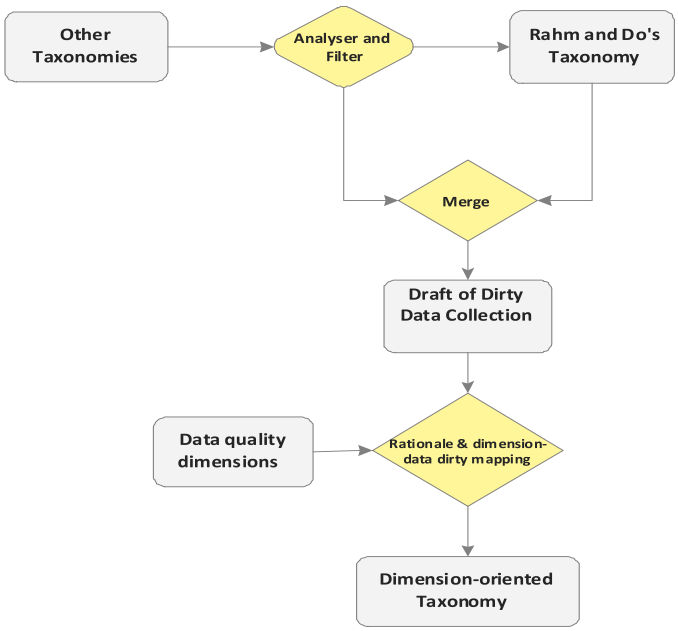

Figure 1. Process of developing an initial taxonomy of dimension-oriented data quality problems

Therefore, we ended up with a draft of types of 'dirty data' that cover all aspects of errors mentioned in the literature earlier. 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 1 shows the initial taxonomy of dimension-oriented of 'dirty data'.

Table 1. The initial taxonomy of dimension-oriented dirty data

ID	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	

4. CONFIRMATION OF THE NEW TAXONOMY

In order to confirm the data quality items associated with data quality dimensions, interviews were conducted with experts and data consumers. Semi-structured interviews were carried out with several experts and EHR stakeholders in the health sector.

4.1 Data Collection

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).

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.[1]

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 2 gives details and justification of the IT experts interviewed in this study.

Table 2. Selected expert interviewees

Participant	Position	Experience(years)	Justification
P1	Database administrator	+15	Direct involvement with quality problems in databases
P2	Database administrator	+10	Direct involvement with quality problems in databases
P3	Physician team leader	+5	Link between medical staff and IT support
P4	Physician team leader	+5	Link between medical staff and IT support
P5	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 those holding academic positions at King Saud bin Abdulaziz University for Health Sciences. Table 3 gives details and justification of the data consumers selected for interview.

Table 3. Selected data consumers for interview

Participant	Position	Experience(years)	Justification
P6	Paediatric consultant	+15	Main decision maker in patients' medical care
P7	Radiology consultant	+10	Data generator and decision maker in patients' medical care
P8	Paediatric consultant + Assistant	+15	Main decision maker in patients' medical care

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

	professor		
P9	Medical director	+10	Decision maker in policy and work regulations
P10	Emergency consultant	+10	Decision maker in a very critical department
P11	Nurse	+13	Data generator due to direct involvement in patient medical care

The interview featured confirmatory and exploratory questions about the data quality problems making up the each dimension. Livescribe[2] pen was used as a tool for recording the interviews.

4.2 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 “overlapped”. The next step was to code and assign data from the transcript to related nodes.

4.3 Findings and Results

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. The results and findings of the interviews were categorised into confirmed; irrelevant; overlapping; and additional items for each dimension.

4.3.1 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 have been through preliminary refactoring and classification, and ended up under the accuracy metric. This would allow objective assessment of the quality of accuracy.

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

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

In addition, an expert added that orphan data could a data quality problem that should be considered within accuracy's.

4.3.2 Consistency

The consistency quality items underwent a process of examination and refactoring to achieve a list of relevant and reliable measures for consistency measurement. Table 4 highlights the outcome of the interviews analysis.

Table 4. Experts' findings for consistency

ID	P1			P2			P3			P4			P5		
	C	I	O	C	I	O	C	I	O	C	I	O	C	I	O
D8	✓			✓			✓			✓			✓		
D9	✓			✓			✓			✓			✓		
D10	✓			✓			✓			✓			✓		
D11	✓			✓			✓			✓			✓		
D12		✓	✓		✓		✓			✓			✓		
D13		✓			✓		✓			✓				✓	
D14	✓					✓	✓			✓			✓		
D15	✓			✓			✓			✓			✓		
D16		✓			✓		✓			✓			✓		
D17		✓			✓		✓			✓			✓		
D18	✓			✓			✓			✓			✓		
D19	✓			✓			✓			✓			✓		

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

The two physician team leaders (P3 and P4) confirmed that all quality items are relevant and sound measures for consistency. However, the two senior DBAs (P1 & P2) and an application analyst (P5) claimed that the item '*referential integrity violation*' is not a consistency-related measure, but is accuracy-related. Moreover, the two DBAs (P1 & P2) considered the items D12, D16 and D17 were sound measures of accuracy. One DBA expert 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 metric items of consistency are comprehensive and cover the required aspects to measure this dimension.

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:

Paediatric consultant: '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. An emergency consultant highlights this by stating:

'We have a noticeable consistency problem with patients' names as one family name could have many spellings in English. Many fatal incidents occurred here [was] caused by such consistency problem.'

4.3.3 Completeness

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

4.3.4 Timeliness

Responses from both the experts and data consumers 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 this metric and the experts did not add any items, believing it was sufficient.

4.4 Discussion

The data quality problems were refactored and mapped against corresponding dimensions. Subsequently, the initial dimension-oriented data quality problems were discussed and validated with two groups, comprised of experts and data consumers. Table shows final model of the dimensions and their compromising data quality problems.

With regard to accuracy-related data quality problems, 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 (P1) 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.

Regarding consistency-associated, 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 'R*eferential integrity violation*' belonged to the accuracy metric, 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 metric.

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.

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

With regard to timeliness-related quality problems, 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-related.

The final result of the proposed taxonomy is displayed in Table 5.

Table 5. The taxonomy of dimension-oriented data quality problems for EHR

Accuracy: The extent to which registered data conforms to its actual value.
1- Illegal values due to invalid domain range
2- Misspellings
3- Misfielded values
4- Embedded values
5- Word transposition
6- Wrong reference
7- Erroneous entry
8- Contradicting records in single/multi source(s)
9- Inconsistent spatial data
10- Referential integrity violation
Consistency: Representation of data values remains the same in multiple data items in multiple locations.
11- Violated attribute dependencies
12- Uniqueness violation
13- Naming conflicts in multi-source
14- Structural conflicts in multi-source
15- Wrong categorical data
16- Duplicated records in single/multi data source(s)

17- Different measure units in single/multi source(s)

18- Syntax inconsistency

19- Inconsistent name spelling

20- Different representations due to use of abbreviation and cryptic values

21- Different representations due to use of Alias/nickname

22- Different representations due to use of encoding format

23- 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

24- Missing data where Null-not-allowed constraint enforced

25- Missing data where Null-not-allowed constraint not enforced

26- Missing record

27- Ambiguous data due to incomplete context

28- Semi-empty tuple

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

29- Outdated temporal value

30- Outdated reference

5. CASE STUDY

Once the dimension-oriented taxonomy of data quality problems and its measures were developed, as part of the study, it is to be validated as a feasibility test for the proposed framework and the defined measurement approach. A case study was needed to evaluate the new approach of assessment using the new taxonomy and to show its practicality and applicability. It was applied in a large-scale hospital to assess and analyse their EHRs to determine quality scores for each dimension. In this hospital, an assessment team was formed of the hospital staff to evaluate their system data and the effectiveness of the new approach.

5.1 Context of Study and Participants

The author conducted the case study in a large hospital serving a population of almost 400,000 patients. The assessment team of 12 people was selected representing all roles in information production, that is, information collectors, information consumers and IT professionals. They are four IT specialists, three people of health informatics department and five consultants and medical staff. Subsequently, they were on an introductory course of how to use the new taxonomy to assess their system data. After that, a sample of 400 patient records was retrieved from their EHR system for the assessment purpose.

5.2 Result of the Data Assessment

The sample of patients' records was taken from three sub-systems of the integrated EHR system, namely SOAP (subjective, objective, assessment, and plan), Laboratory and Pharmacy systems. The Figure 2 shows the result of the assessment process. LAB and Pharmacy systems scored almost 100% quality of data in all dimensions. It was clearly observed that these two systems are having 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. This may result in very good quality of data in the two systems.

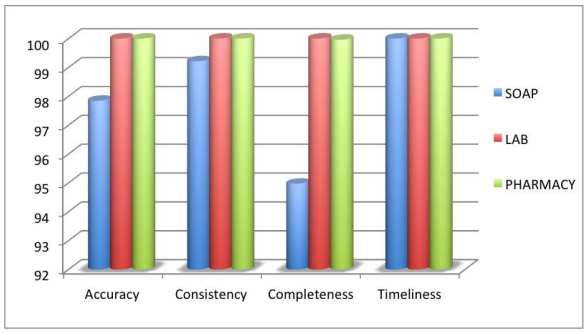

Figure 2. Assessment scores of the three systems

On the other hand, SOAP system reported a relatively lower score in the data quality comparing to the other systems. The data quality constraints and precautions for this system need reconsideration. 'Misspellings' and 'illegal values due to invalid domain range' measures affected the accuracy of the patients' data populating the SOAP system, while 'semi-empty tuples' measure brought down the quality level of the data completeness in the system.

In general, the EHR system scored a very good level of quality in accuracy, consistency, completeness and timeliness dimensions using the proposed taxonomy. This new approach highlighted some issues of the SOAP system in terms of accuracy and completeness.

5.3 Validation

The next step was to evaluate the new taxonomy. Ten members of the assessment team were asked to evaluate the effectiveness of the taxonomy through a questionnaire developed by the researcher. The questionnaire examined 'the ease of the use', 'its perceived usefulness', 'the user satisfaction' and 'perception of congruence between expectation of the use and its actual performance' of the proposed approach.

The questionnaire data were analysed using SPSS software to examine the perception of the assessment team towards the taxonomy. The hypothesis was tested using One-sample T test.

Table 6. One-sample statistics for the results of taxonomy evaluation

	N	Mean	Std. Deviation	Sig. (2-
Perceived ease of use				
Learning to operate **Data Quality Assessment tool** is easy for me	10	4.1	0.56	<0.01
I find it easy to get **Data Quality Assessment tool** to do what I want	10	4.1	0.73	<0.01
It is easy to become skillful at using **Data Quality Assessment tool**	10	4.1	0.73	<0.01
Overall, I find **Data Quality Assessment tool** easy to use	10	4.4	0.51	<0.01
Satisfaction				
I am satisfied about the quality results I got after using the tool	10	4.5	0.52	<0.01
I am pleased for the overall quality of our data after using the tool	10	4.2	0.63	<0.01
I am content with the experience of using the tool	10	4.0	0.66	<0.01
How would you rate your overall satisfaction with us?	10	4.3	0.48	<0.01
Perceived usefulness				
Using this tool helps assess the quality of our system data.	10	4.0	0.66	<0.01
Using this tool increases my productivity in assessing and measuring the quality of our medical data.	10	4.2	0.78	<0.01
Using this tool enhances my effectiveness in managing and assessing the quality of our medical data.	10	4.2	0.63	<0.01
Overall, this tool is useful in assessing and assuring the quality of our medical data.	10	4.1	0.56	<0.01
Confirmation				
My experience with using this tool was better than what I expected.	10	4.0	0.66	<0.01
The service level provided by this tool was better than what I expected.	10	4.1	0.56	<0.01

As shown in Table 6, the analysis results show that participants agreed on the effectiveness of the proposed approach. The fact that all answers were significant, as *p* values for all dimensions were less than 0.05, confirms that participants' perception were significantly positive towards the new taxonomy. Figure 3 implies that they found the tool is easy to use and follow to conduct data quality activities. It also indicates that they perceived the usefulness of the proposed approach, and satisfied of its results.

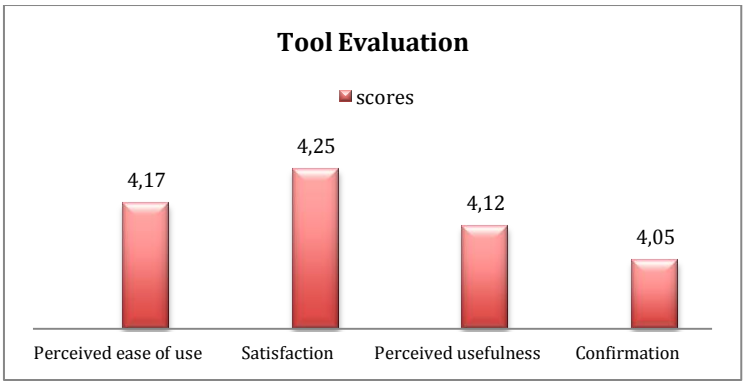

Figure 3. The mean of each scale of the tool evaluation

Figure 4 depicts the different perceptions of the information production roles towards assessing the data using the new taxonomy. All roles (IT staff, health informatics and consultants) perceive almost the same level of usefulness of the new approach achieving 4.06, 4.17 and 4.17 respectively. However, IT staff scores the lowest in the scale of the 'perceived ease of use'. This could be that the technical tasks of data quality assessment assigned to them were time-consuming and needed more effort comparing to the tasks assigned to the other roles. With regard to the satisfaction, all roles were satisfied with results but the health informatics, who are responsible for maintain the patients records, scored higher.

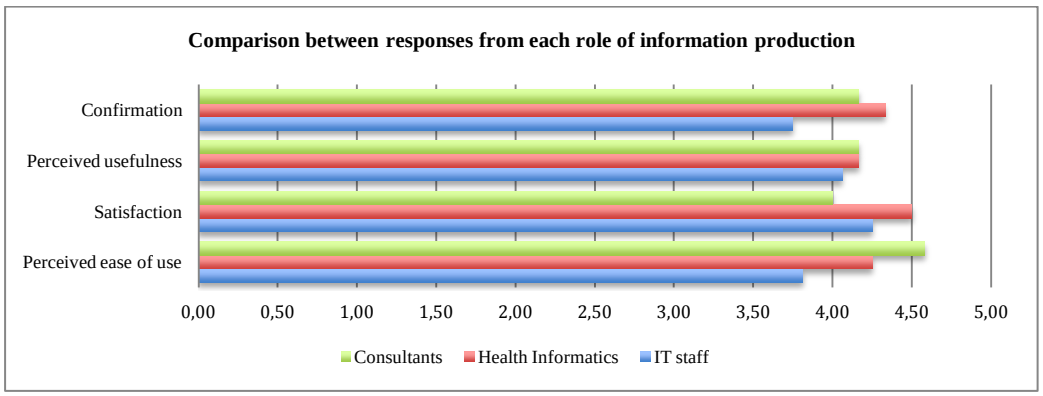

Figure 4. Mean comparison among roles in information production

6. CONCLUSION AND FUTURE WORK

In this paper, the 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. Thus, the proposed taxonomy concerned with identifying the problems from the perspective of quality dimensions. This will help health organisations

prioritise the data quality problems associated with most desirable dimensions in the process of data quality assessment. Such 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.

Future work involves the development and deployment of severity factors that make quality problems severer into these data quality problems. These factors were clearly noticed during the interviews. Such mechanism would help organisations save time and money by prioritising the severest problems.

This study is ethically approved by the The University of Southampton Ethics Committee (Submission ID: **16276**).

REFERENCES

Almutiry, O., Wills, G. & Crowder, R., 2013. Toward A Framework For Data Quality In Electronic Health Record. In *e-Society conference*.

Aronson, J., 1994. A pragmatic view of thematic analysis. *The qualitative report*, 2(1), pp.1–3.

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

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. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041534&tool=pmcentrez&rendertype=abstract.

Braun, V. & Clarke, V., 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), pp.77–101. Available at: http://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

Canadian Institute for Health Information, 2009. *The CIHI Data Quality Framework*,

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 June 1, 2014].

Ge, M. & Helfert, M., 2008. Data and information quality assessment in information manufacturing systems. *Business Information Systems*. Available at: http://link.springer.com/chapter/10.1007/978-3-540-79396-0_33 [Accessed June 12, 2014].

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*,

Kim, W. et al., 2003. A taxonomy of dirty data. *Data mining and knowledge …*. Available at: http://link.springer.com/article/10.1023/A:1021564703268 [Accessed June 7, 2014].

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.

Müller, H. & Freytag, J.-C., 2005. *Problems, methods, and challenges in comprehensive data cleansing*, Professoren des Inst. F{ü}r Informatik.

Oliveira, P. & Rodrigues, F., 2005. A taxonomy of data quality problems. *... and Information Quality ...*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.109.1695&rep=rep1&type=pdf [Accessed June 7, 2014].

Rahm, E. & Do, H.H., 2000. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4), pp.3–13.

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.

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. *J. of Management Information ...*, 12(4), pp.5–33. Available at: http://courses.washington.edu/geog482/resource/14_Beyond_Accuracy.pdf [Accessed May 27, 2014].

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.

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.