Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-721

Organizing Data Quality Assessment of Shifting Biomedical Data

Carlos SÁEZ[a,1], Juan MARTÍNEZ-MIRANDA[a], Montserrat ROBLES[a] and
Juan Miguel GARCÍA-GÓMEZ[a]
[a] *Grupo de Informática Biomédica (IBIME), Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Valencia, Spain*

Abstract. Low biomedical Data Quality (DQ) leads into poor decisions which may affect the care process or the result of evidence-based studies. Most of the current approaches for DQ leave unattended the shifting behaviour of data underlying concepts and its relation to DQ. There is also no agreement on a common set of DQ dimensions and how they interact and relate to these shifts. In this paper we propose an organization of biomedical DQ assessment based on these concepts, identifying characteristics and requirements which will facilitate future research. As a result, we define the Data Quality Vector compiling a unified set of DQ dimensions (completeness, consistency, duplicity, correctness, timeliness, spatial stability, contextualization, predictive value and reliability), as the foundations to the further development of DQ assessment algorithms and platforms.

Keywords. data quality, decision-making, electronic health records, dataset shifts

1. Introduction

The lack of Data Quality (DQ) is an important open issue that leads into poor decisions and suboptimal processes. This is particularly important in the healthcare information, where the quality of data may have direct consequences on the care process of the patients. This may lead physicians to a set of direct errors, such as inappropriate or outmoded therapy, technical surgical error, inappropriate medication, error in dose or use of medications; and indirect errors, such as failure to take precautions, failure to use indicated tests, avoidable delay in diagnosis, failure to act on results of tests or findings, and inadequate follow up of therapy [1].

In addition, insufficient DQ may directly harm the results of studies that re-use the data, such as clinical trials or cohorts. Much of the limitations to exploit the clinical information are related with the fact that the original Electronic Health Records (EHRs) are designed for a restricted primary purpose, but without taking into account secondary use of data that may require different levels of quality [3].

Up to date, several approaches have been proposed to organize the concepts associated to DQ from different perspectives, such as by DQ dimensions [8, 16], processes [8, 10], or requirements [2, 5]. Others focused their studies in analysing the quality of biomedical data, most based in the measurement of DQ dimensions,

[1] Corresponding Author: Carlos Sáez (carsaesi@ibime.upv.es), Grupo IBIME, Instituto ITACA, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022, Valencia, Spain

assuming stationary distributions in the datasets, or regarding to the use of clinical information standards. These approaches miss the shifting characteristics related to the changes in data underlying concepts caused, as introduced in [3], when collecting data for long periods of time and, in our opinion, also when dealing with multi-centre or multi-user data. This effect, related to non-stationary distributions, can be studied under the problem of dataset shift [9, 14].

Nevertheless up to date, we have not found any approach for classifying the concepts of DQ assessment for biomedical environments that entails the aforementioned data shifting-related characteristics. Thus, we think that a novel unified organization of DQ assessment for biomedical data (i.e. compiling shifting-related characteristics of DQ, functional requirements and outcomes) will be helpful for a holistic treatment of the problem in future works on the discipline. Particularly, in this work we introduce and discuss the general foundations of a Data Quality Vector (DQV), as an adaptable framework handling a complete set of DQ dimensions to facilitate DQ assessment of biomedical data, considering the aforementioned characteristics.

2. State-of-the-art

Data can be considered a product manufactured by organizations [13]. Under this assumption, the Massachusetts Institute of Technology (MIT) launched in 1992 the Total Data Quality Management (TDQM) program [12], based in the features of Total Quality Management introduced in early 1980's for the management of quality in industry. The TDQM cycle suggests continuous improvement of data quality based on 4 stages: 1) define, 2) measure, 3) analyse and 4) improve. TDQM also applies to biomedical data. Even though biomedical data in most cases represent a patient's status, data itself is produced by medical staff as well as by devices. This batch DQ control is e.g. well-established in clinical laboratories by means of Levey-Jennings charts and Westgard rules.

Quality of biomedical data has been principally studied in data repositories for study cohorts [4, 15] and for the integration of heterogeneous sources [3]. DQ of routine EHRs has also been studied [3, 11]. Most of these studies focused to the measurement of DQ dimensions and to the use of biomedical information standards. However, the association of the shifting-related characteristics of DQ, mainly related to dataset shifts, is an open research topic.

There is a general agreement about an initial phase to define DQ in terms of fitness for use [8, 13] or, as proposed in [7], as data quality goals. This customization of DQ to specific purposes can be based e.g. on a set of constraints over DQ dimensions. Many studies defined some DQ dimensions [3, 8, 16]; however, we still find some gaps and discordances among them.

Recently, in [5], after a review work of industrial tools for quality assessment, Gartner classified the core functional requirements of the DQ discipline in: profiling, parsing and standardization, cleansing, matching, monitoring, and enrichment. In this work we additionally provide a classification of functions and outcomes of DQ analysis specific for biomedical data, considering the aforementioned shifting-related characteristics.

3. Organizing data quality assessment of shifting biomedical data

We have identified the characteristics, functionalities and outcomes that characterize the data quality assessment considering the shifting behaviour of data. Table 1 presents a classification of the shifting-related characteristics of the data relevant for the DQ analysis. These characteristics are associated to the production of new data in the curse of time and across space or populations. Table 2 classifies the temporal characteristics that can be considered in the DQ analysis procedure. The table 3 classifies the expected functionality of a DQ analysis system for biomedical data. Finally, Table 4 classifies the expected outcomes of such systems.

Table 1. Shifting-related data characteristics for DQ analysis

Characteristic	Classification	Description
Time	Time-stamped	Data has information about acquisition time
	Non time-stamped	Data does not have information about acquisition time
Inter-population	Local	Data created by single centre
	Multi-user	Data created by multiple users
	Multi-centre	Data created by multiple centres

Table 2. Temporal characteristics of DQ analysis

Characteristic	Classification	Description
Time dependency	Time dependent	The DQ analysis considers a temporal relation among data
	Time independent	The DQ analysis does not take into account any temporal relations
Data gathering	On-line	Data is gathered as a continuous flow (or data-stream)
	Off-line	Data is gathered as a dataset
Time constraints	Reactive or time constrained	The DQ analysis must provide a result before a specified time
	Non time-constrained	There are not restrictions in execution time for the DQ analysis
Period	Short-term	The DQ analysis is performed in data acquired during the current period of time
	Long-term	The DQ analysis looks for DQ problems in data acquired along a time period

Table 3. Functionalities of DQ assessment for biomedical data

Functions	Description
Single case quality assessment	DQ analysis for a single case in insert, update or retrieval time
Continuous DQ monitoring	A monitor of the DQ of streams or batches
Alerts about DQ	The system triggers an alert when a predefined DQ goal is not achieved
Data selection	The user wants to obtain a set of data that fulfils a set of DQ requirements
Generate DQ Reports	Obtain a DQ report based in a predefined or custom query
Data integration	Control DQ in the integration of data in a centralized or federated database

Table 4. Outcomes of DQ assessment for biomedical data

Classification	Description
DQ levels	The measurements of DQ dimensions or functions of them (see Table 5)
Set of high-quality data	A set of data that fulfils some DQ requirements
Track of low DQ causes	Hints for the possible causes of recurrent low DQ
Trends of DQ	An analysis of trends of DQ

It is straightforward to see that most classified concepts can be related among them, inter and intra-table. For instance, a DQ monitoring system can check DQ alerts based on dataset shifts, thus data must be time-stamped and generated on-line, the DQ analysis will be time-dependent for short-term data and the outcome can be a function of DQ levels. Tables 1 and 2 may facilitate the association of the shifting behaviour of the data in real scenarios with the DQ analysis to implement the functions of Table 3 and achieve the outcomes of Table 4.

A good scenario where the proposed integrated organization can be applied is in massive-data environments, such as a regional or national Healthcare service.

4. Unifying concepts in a Data Quality Vector

As a consequence of the introduced organization, we propose the Data Quality Vector (DQV) as a holistic view of the Biomedical Data Quality. It is intended to establish the foundations for development of general DQ metrics and algorithms, particularly those envisaging the shifting-related characteristics of DQ analysis. Based on the literature and the organization presented in section 3, the DQV includes the nine DQ dimensions presented in Table 5 defined with the purpose to cover all the previously proposed dimensions in the literature. The DQV intends to measure such DQ dimensions independently or as function of them. Complementary to [7, 4] we propose that DQ goals can be customized based in a function of a combination or a set of constraints among the DQ dimensions.

Currently, our efforts focus on defining the metrics for each dimension, where, according to the purpose, some of them will be classified as generic (i.e. domain-independent, such as a degree of the number of duplicated data) and some others as domain dependent (parameterized given a scenario, such as measuring the predictive value for a specific decision support task). Additionally, data stream-mining, scalable learning, and reactive algorithms are being studied to implement the functionalities described in Table 3 for shifting data. Special emphasis will be put to analyse the interactions among short and long-term DQ analysis, based in the changes or recurrences of the underlying data concepts through time or across populations.

Table 5. DQ dimensions in the DQV

Dimension	Description
Completeness	The degree to which relevant data is recorded
Consistency	The degree to which data satisfies specified constraints and rules
Duplicity	The degree to which data contains duplicate registries representing the same entity
Correctness	The degree of accuracy and precision where data is represented with respect to its real-world state
Timeliness	The degree of temporal stability of the data
Spatial stability	The degree to which data is stable among different populations
Contextualization	The degree to which data is correctly/optimally annotated with the context in with it was acquired
Predictive value	The degree to which data contains proper information for specific decision making purposes
Reliability	The degree of reputation of the stakeholders and institutions involved in the acquisition of data

5. Conclusion

Dynamic features in DQ analysis, which includes data shifting, were already stated in [2, 3] as an open research topic in DQ. In this work we have proposed an organization of characteristics, functionalities and outcomes of DQ assessment associated to the shifting-related properties of data. As a result, we have introduced the general bases of the DQV, the first step of an on-going work aimed to establish the foundations for the further development of relevant DQ metrics, algorithms and tools.

Regarding to the industry of DQ tools, we complement the functional requirements defined by Gartner [5] with specific requirements for biomedical DQ assessment.

According to the TDQM cycle the DQV will cover stage 1: the DQV is adaptable to the domain; stage 2: the DQV is intended to measure DQ; stage 3: the customized functions of DQ facilitate the posterior analysis; and stage 3: through the algorithms associated with the corresponding items in Tables 2. In general the DQV contributes to improve the DQ, as defined in stage 4.

We are currently working in completing the theoretical foundations of the DQV as well as in defining a robust evaluation framework for its methods, while improving the following up research on dataset shifts associated to DQ. In the future work, we will also study the further inclusion of the ISO 8000 standard (currently under development and with the objective to assess organizations on meeting data quality requirements) which might help in the complete definition and the usability of the DQV.

References

[1] Aspden P et al. Patient Safety: Achieving a New Standard for Care. Committee on Data Standards for Patient Safety. The National Academies Press, Washington, D.C. 2004. ISBN 0-309-09077-6

[2] Berti-Équille L and Dasu T. Data Quality Mining: New Research Directions. ICDM 2009, Miami

[3] Cruz-Correia RJ et al. Data Quality and Integration Issues in Electronic Health Records. V. Hristidis (ed.) Information Discovery On Electronic Health Records. 2010. 55-96

[4] Etcheverry L et al. Data Quality Metrics for Genome Wide Association Studies. 2010 Workshops on Database and Expert Systems Applications

[5] Friedman T and Bitterer A. Magic Quadrant for Data Quality Tools. Gartner Research Note G00214013. 2011 Jul

[6] German RR et al. Quality of cancer registry data: findings from CDC-NPCR's Breast and Prostate Cancer Data Quality and Patterns of Care Study. J. Registry Manag. 38(2): 75-86. 2011

[7] Jeusfeld MA et al. Design and Analysis of Quality Information for Data Warehouses. Proc of the 17th International Conference on Conceptual Modelling; 1998; 1:349:362

[8] Karr AF et al. Data quality: A statistical perspective. Stat Meth; 2006; 3:137-173

[9] Klinkenberg R. Learning drifting concepts: example selection vs. example weighting. Intell Data Anal; 2004; 8(3):281:300

[10] Lee YW et al. AIMQ: a methodology for information quality assessment. Inf and Manag; 2002; 40:133-146

[11] Liaw ST. Data quality and fitness for purpose of routinely collected data – a general practice case study from an electronic Practice-Based Research Network (ePBRN). AMIA Annu Symp Proc. 2011:785-794

[12] Madnick SE and Wang RY. Introduction to total data quality management (TDQM) research program. TDQM-92-01, Total Data Quality Management Program, MIT Sloan School of Management. 1992

[13] Madnick SE et al. Overview and Framework for Data and Information Quality Research. ACM J Data Inform Quality; 2009; 1 (1), Article 2

[14] Moreno-Torres JG et al. A unifying view on dataset shift in classification. Pattern Recogn; 2012; 45:521-530

[15] Müller H and Naumann F. Data Quality in Genome Databases. International Conference on Information Quality. 2003:269-284

[16] Wang RY and Strong DM Beyond Accuracy: What Data Quality Means to Data Consumers. J Manag Inform Syst; 1996; 12 (4):5-34