Addressing Data Quality in Healthcare

Kalinka Kaloyanova[1,2], Ina Naydenova[2], Zlatinka Kovacheva[2,3]

[1]Faculty of Mathematics and Informatics, Sofia University St.Kliment Ohridski
5 James Bourchier blvd., 1164, Sofia, Bulgaria
[2]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
[3]University of Mining and Geology „St. Ivan Rilski“

kkaloyanova@fmi.uni-sofia.bg

Abstract. Data quality is an important part of information processing, but its application in practice is often underestimated. The complexity of data quality management, especially in the case of big data, makes it difficult to work in different areas of application. Although medical records are a significant source of errors in most cases data quality assessment on medical data is partially performed. The presented data quality analysis and recommendations in this paper can help physicians and software developers to understand better data quality dimensions, identify gaps in quality assessment, and develop |own procedures and techniques that correspond to their specific use cases.

Keywords: Data Quality, Data Quality Dimensions, Healthcare, Medical Records.

1 Introduction

The use of software applications in healthcare is growing constantly. Data stored in various information systems help medical workers to provide efficient treatment of their patients every day. Moreover, this data can also be used for statistics, analysis, prognoses. The secondary use of clinical data has been established in recent years as a promising direction for data analysis and decision-making in healthcare. It can be used to optimize the workflow in hospitals and other medical centers.

The quality of clinical data has not only an immediate impact on operative medical processes but also a long-term influence on the research of accumulated and aggregated clinical data. Inconsistency in data, missing values, invalid data – all forms of uncertain or inaccurate data, questions the usefulness of the collected data. Ignoring data quality issues leads to the worthlessness of data collection efforts because no value is created.

Big data raises new issues in data management at all, including the data quality processes. New dimensions of data quality should be addressed, to overcome these challenges of the huge amount of data coming from different sources, stored in different ways (structured and unstructured data), the speed of data generation and how to handle it in a timely manner.

In this paper we discuss important aspects of data quality, refer to the common data quality dimensions and analyze their application in the healthcare domain. Finally, we present a set of guidelines for the implementation of a successful data quality process.

2 Common data quality dimensions

First researches on data quality appeared in 1990 and different definitions were proposed during the years. Later, the main principles of data quality were described in the standard ISO/TS 8000-1:2011 [5]. ISO/IEC 25000:2014 defines *data quality* as "degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions"[6].

To understand data quality, different data quality dimensions should be considered. Dimensions represent measurable data quality characteristics. Some of these dimensions are related to the particular domain, users, or services. ISO/IEC 25012 focuses on Data Quality Model and classifies 15 data quality characteristics into two groups: Inherent Data Quality and System Dependent Data Quality [7]. The group of inherent characteristics presents dimensions that are relevant in most cases such as accuracy, completeness, and consistency, currentness (timeliness):

- *Accuracy* – presents the degree to which attributes of data correctly represent the true value of the characteristics of the intended object. Accuracy can be seen as syntactic as well as semantic accuracy.
- *Completeness* – measures the degree to which an entity has values for all attributes that are expected.
- *Consistency* – the degree to which data attributes have no conflict and are consistent with other data (and their attributes).
- *Currentness* – reflects how sufficiently up-to-day are data attributes in the specific context of use.

Accessibility, creditability, compliance, efficiency, and *confidentiality* complement this group. The other group – System Dependent Data Quality group consists of dimensions, that reflect the degree to which the quality of data in a computer system is achieved and maintained when the data is used under certain conditions. This includes *availability, recoverability, portability*, as well as *precision, traceability,* and *understandability* [6].

Wang in [17] and [18] commenced in-depth research on data quality and its dimensions more than twenty years ago. Further, many researchers discuss data quality problems and characteristics [13], [15], [16]. A survey, presented in [12] observes a variety of sources and discusses data quality dimensions and identifies the most frequently cited of them. It also concerns poor data classification issues.

Recently, data quality understanding evolves in the case of Big data [1], [4]. Not only the enormous volumes of data are the challenge, but also the diversity of their sources and applications domains – documents, images, audio, video, maps, linked open data, social media, sensor data [3]. So a new dimension *trust* that reflects the reputation and reliability of the source is also considered [2]. Data types also vary and the part of unstructured data and semi-structured data is growing more and more than the structured ones.

Rapid changes in data values raise the question of their *timeliness*, which is too short in many cases. This, in connection to data *veracity*, raises the question of the objectivity of the data and therefore of the analyses that are made on them. The *validity* of data should be at the highest level, too, in order to bring a business value [4].

3 Data quality challenges in the eHealth area

Despite common problems with data quality, specific domains also influence quality dimensions. Health information is used to aim medical decisions for primary patient care and to support the continuity of treatment between different medical providers [2]. It is crucial for these decisions to be made on the basis of complete and reliable data.

Medical data provide knowledge about individual patients or groups of patients – facts about their health condition that are subject to further processing. These are mostly *textual data and numeric values but also can include images, audio, and even video data.*

Individual data present the health characteristics of individual patients, most often in the context of a disease. Both, structured and unstructured data are used. Data quality dimensions *accuracy, completeness, correctness* are of great importance here. In many cases, free text data fields are used in primary care software systems that do not require all needed information to be entered and enable this information to be presented in a non-consistent way. There are cases where some data fields in medical records stay empty as the requested information is not entered or is not yet known. As doctors have to balance between patient care and data entering, the information side is not the priority. Measurement errors, recording errors, or transcription mistakes are often seen [9]. Other errors arise when documents are transferred between different systems.

An essential aspect of medical data is its *confidentiality* – who are the authorized persons that have access to the particular piece of information and in which cases they could use it [2].

In addition to the primary use of health information, the obtained data can be summarized and further processed to serve for more in-depth research, statistics, trends detection. In the case of secondary use of clinical data, the *volatility* dimension, which reflects how quickly data is changed, becomes important. Invalid or inaccurate data also may influence the obtained conclusion. Other quality dimensions such as *usability, usefulness,* and *relevance* complement the main set of data quality characteristics *accuracy, validity, completeness*, and *currentness***.**

Using standards is a common way to overcome many issues in data processing. Health Level 7 (HL7), ISO/IEC 13606, Systematized Nomenclature of Medicine (SNOMED), Digital Imaging and Communications in Medicine (DICOM) have been established as basic international standards for health data processing [2]. But their implementation at all levels of data processing is not yet widespread. Semantic interoperability is still a challenge for software applications presenting medical data in most countries. Other standards in connection with data presentation in software systems are noted in [11], [9]. The use of medical devices sending streaming information poses other challenges inherent in Big data [15]. Even the use of the appropriate standards is not able to guarantee full prevention from data errors because they evolve constantly [10].

Data codification is another approach to overcome the incompatibility of data and to achieve unification. The International Classification of Diseases (ICD) goes through several versions to reach ICD-11 revision. The Anatomical Therapeutic Chemical (ATC) classification system classifies the active ingredients of drugs and it is used as a pharmaceutical coding system [2].

4 Motivation case study

Various computer applications operate in Bulgarian healthcare. They generate a significant amount of data. Medical data is often presented in XML format. In Bulgarian healthcare, the National Health Insurance Fund (NHIF) also uses this approach. A set of predefined XSD schemes provides various templates for different medical providers in the country in order to achieve regular information about their activities to NHIF. The templates consist of predefined data fields presenting common information about doctors and medical practice, patients' personal information, as well as information about their medical condition and treatment.

As a motivation case study, we considered a multitude of samples obtained from different software applications used by general practitioners, hospitals, and several other medical centers in the country. All they present information in XML format, following the XSD schemes, provided by the NHIF site for the particular medical providers [14].

In addition to the administrative information for patients and medical institutions, these records contain information about patients' illnesses, medications used, treatment periods, and other data, specific to medical provider and application type.

For example, the Outpatient card – one of the main documents, used by general practitioners in Bulgarian healthcare, presents information about patients visits. Every year more than 25 000 000 patients' visits are recorded using this template [8]. Some of the sections of this template allow free text entry and the use of abbreviations and acronyms that significantly complicates *understanding* and further processing of this information. In this way data *accuracy, integrity* and *completeness* are not met or partially implemented.

Information systems used in hospitals in the country also allow such free data fields. We obtained and analyzed XML extracts with information about thousands of patients of a hospital, many drug protocols, clinical procedures, dispensary observations, etc. All personal data was anonymized. In the case of hospitals, inconsistency in the date *in* and *out* of a patient in the hospital is a potential source of erroneous data, as there exist different types of XML extracts for patients, admitted to the hospital and the leaving ones. This additionally affects the data *consistency* dimension.

Possible conflicts of information could be found even in more structured data fields, which are usually used for administrative information. For example, when comparing gender field and information, extracted from patient EGN (Personal identification number in Bulgaria). Mismatched numbers of medical practices, branch numbering, and other demographic data fields with data out of range.

Further, medical data could be incorrectly recorded. For example, the results of a laboratory test or blood pressure measurement may be partially or completely wrong, which raises again *accuracy* and *validity* problems. Moreover, there could be *incompatibility* on a logical level – between patient's diagnosis and prescribed medicine or between patient's age and medical procedures or prescribed medicine.

Taking into account the number of medical records, produced from the various systems, data *usefulness* should be considering be when data is collected. This will reflect on the cost for assembling, storing, processing, and dissemination of the information. The right balance between all dimensions will improve the performance of all data management activities.

5 Guidelines for data quality process implementation

A primary goal for establishing an efficient framework for data quality management is to prevent data errors in collecting and storing data. Based on the issues discussed above we outline a set of guidelines that could be used to set up a framework for data quality prevention and management in Bulgarian healthcare.

5.1 Quality planning

The first step for building a common framework for data quality management requires a systematic way to cover all aspects concerning data quality. Data quality planning will help organizations to maintain the balance between data quality goals and resources needed to reach them. To be a successful one, it should include:

- Identifying all roles, involved in the data quality management process

First, these are medical workers – doctors, nurses, therapists, dentists, pharmacists, etc., insurers, representative of administrative structures in healthcare at a local and national level. Also, policymakers and lawmakers. Patients and their relatives should be considered, too.

- Identifying data to be collected

All data sources should be identified, the user's requirements about data – collected and structured, and data formats – defined.

- Determining data quality goals and dimensions

The list of quality characteristics should be established following common data quality dimensions and choosing additional dimensions that are most appropriate to reach the goal of the particular case.

- Choosing appropriate standards

The set of used standards depends on data sources and data formats and the goals and requirements for the system to be implemented.

- Defining rules

Different rules can be considered for incompatibility, incompleteness, duplication of data. Other groups of rules can concern values out of range, temporal sequence errors, data that is incompatible with other data.

- Defining metrics

The state of data can be measured based on standard metrics, applicable for every particular data set. In addition, specific to the area and application context business constraints could be defined.

5.2 Quality assurance

Besides standard recommendations for quality assurance, it will be useful as much as possible data errors' eventualities to be prevented through implementation in the software applications. Below several suggestions for the development of software systems in healthcare are listed:

- Applying data patterns:
  - Use of standards;
  - Use of archetypes;
  - Use of international coding conventions.

The application of specific to healthcare standards will facilitate data processing. Previously defined archetypes can control values, that describe different parameters of patient' state. The implementation of the coding conventions like IDC and ATC allow storing the codes into appropriate structures and choosing, instead of entering values, which will prevent typing errors. It also supports the semantic interoperability between different systems.

- Standardization of data entering processes
  - Use of appropriate user interface;
  - Use of data entry templates;
  - Validation of data entry fields.

The use of common interface templates helps users to feel confident when working with the system. User control and freedom should be provided consistently through all system functions. The timely verification of the entered data will avoid storing wrong data into data structures.

- Implementing rules

Software implementation of all rules, defined during Quality planning is an efficient way to prevent data errors. Appropriate error messages should be considered when rules are broken.

- Checking data using software tools
  - Providing data entering fields based on standards;
  - Validation of the range of parameters;
  - Checking diagnoses, and drug compatibility;
  - Writing special application programs.

All data should be validated on input. Recognition of data (using predefined values, codes, archetypes, etc.) instead of direct data entering, will avoid entering incorrect data.

- Providing appropriate kinds of help, documentation, and training.

In many cases training of different groups of users is recommended – for example, administrative staff, clinical staff, etc. Well-structured documentation, help pages, and other supporting documentation have proven their effectiveness.

5.3 Quality control

Data Quality Control follows all recommendations of the plans established during Quality Planning. Data should be processed according to the specified rules, following the prescribed procedures. Taking appropriate feedback, providing regular quality reviews, will help to identify and appropriately react every time when data processing does not meet data quality requirements.

In order to provide a common language and a harmonized approach to measuring and improving data quality in the eHealth area, the World Health Organization in collaboration with other organizations proposes a Data Quality Review (DQR) toolkit and methodology. It includes guidelines that promote framework and institutionalization of regular and independent review and assessment of the data quality state at different levels of the health system in countries (national, district, and facility). Several tools that can be adapted by users are also included [19].

5.4 Quality improvement

Considering the results coming from the data quality monitoring and control activities (system assessment, data verification, help desk reviews, etc.) it is reasonable for a working group on data quality to lead the development and implementation of a data quality improvement plan. It is recommended, when working on the plan to:

- outline the activities, that will address the problems, pointed out during the assessment;
- allocate the resources;
- identify the staff, that will provide all procedures to improve the quality of data.

As for the list of the activities, first should be implemented these ones, which will cause a major impact on overall data quality [20].

6 Conclusions

In this paper important characteristics of data quality were discussed and related to data used in healthcare. The quality of data is essential to achieve the full potential of healthcare data accumulated so far and the quality characteristics of this data should match certain levels. We analyzed major data quality issues and formulated a set of guidelines for data quality planning, assurance, and control that could be successfully applied for the healthcare domain.

All presented guidelines are subject to many extensions especially in terms of their practical implementation. Considering the significant role of the specific subject area, defining a data quality assessment model for medical data used in Bulgarian healthcare is one of our next goals.

7 Acknowledgments

This paper is supported by the project BG05M2OP001-1.001-0004 Universities for Science, Informatics and Technologies in the e-Society (UNITe) and the National Scientific Program “eHealth” in Bulgaria.

References

1. Batini C., Rula A., Scannapieco M., and Viscusi G.: From Data Quality to Big Data Quality. J. Database Management, vol 26, no 1: 60–82 (2015).
2. Batini C., and Scannapieco M.: Data and Information Quality, Springer (2016).

3. Cai L., Zhu Y.: The Challenges of Data Quality and Data Quality Assessment in the Big Data Era Data Science Journal, 14: 2, pp. 1-10 (2019).
4. Caballero I., Serrano M., Piattini M.: A Data Quality in Use Model for Big Data. In Advances in Conceptual Modeling. ER 2014. Lecture Notes in Computer Science, vol. 8823: 65-74. Springer, Cham (2014).
5. ISO/TS 8000-1:2011, Data quality – Part 1: Overview https://www.iso.org/standard/50798.html, last accessed 2021/03/12.
6. ISO/IEC 25000:2014, Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaREISO/IEC, "Software engineering – Software product Quality Requirements, https://www.iso.org/obp/ui/#iso:std:iso-iec:25000:ed-2:v1:en, last accessed 2021/03/18.
7. ISO/IEC 25012 Software and Data Quality, https://iso25000.com/index.php/en/iso-25000-standards/iso-25012, last accessed 2021/03/18.
8. Kaloyanova K., Krastev E. and Mitreva E.: Extracting Data from General Practitioners' XML Reports in Bulgarian Healthcare to Comply with ISO/EN 13606. In: Proceedings of the 9th Balkan Conference on Informatics (BCI'19), Sofia, Bulgaria, ACM DL, Article 3 (2019).
9. Kim F.: Data Quality in Healthcare – Challenges, Limitations & Steps to Take for Quality Improvement, https://dataladder.com/data-quality-in-healthcare-data-systems, last accessed 2021/03/12.
10. Krastev E., Tcharaktchiev D., Kirov L., Kovatchev P., Abanos S. and Lambova A. (2019) Software Implementation of the EU Patient Summary with Archetype Concept, In: Proceedings of GLOBAL HEALTH 2019, The Eighth International Conference on Global Health Challenges, Porto, Portugal, September 22-26, pp. 8-13 (2019).
11. Krastev E., Tcharaktchiev D., Kaloyanova K., Kirov L., Kovatchev P., Abanos S., Mateva N. (2020) Standards Based Adaptation of Clinical Documents for Interoperability of e-Health Services, In: Proceedings of the 13th Conference on Information Systems and Grid Technologies (ISGT 2020), Sofia, Bulgaria, May 29 – 30, 2020, CEUR-WS.org/Vol-2656/paper2.pdf, last accessed 2021/03/17.
12. Laranjeiro N., Soydemir S. N., Bernardino J.: A Survey on Data Quality: Classifying Poor Data, In: Proceedings of the IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC), Zhangjiajie, China, pp.179-188 (2015).
13. Mahn-DiNikola V.: Six Dimensions of Data Fitness, https://blog.medisolv.com/articles/six-dimensions-of-data-fitness, last accessed 2021/03/31.
14. NHIF. National Health Insurance Fund. "XML file format for submitting requests by doctors and specialists for accounting completed ambulatory activities in primary and specialized patient care after January 1st., 2021", https://www.nhif.bg/page/28, last accessed 2021/02/15.
15. Richardson I.: Healthcare Systems Quality: Development and Use, In: Proceedings of the International Workshop on Software Engineering in Healthcare Systems, Austin, USA, pp. 50–53 (2016).
16. Ristevski B., Savoska S., Blazheska-Tabakovska N.: Opportunities for Big Data Analytics in Healthcare Information Systems Development for Decision Support, In: Proceedings of the 13th Conference on Information Systems and Grid Technologies (ISGT 2020), Sofia, Bulgaria, 2020, CEUR-WS.org/Vol-2656/paper4.pdf, last accessed 2021/03/12.
17. Wang R. Y. and Strong D. M.: Beyond accuracy: What data quality means to data consumers, Journal of Management Information Systems, Vol. 12, no. 4: 5–33 (1996).
18. Wang, R.Y.: A product perspective on total data quality management, Communication of ACM Vol 41, no 2: 58–65 (1998).
19. World Health Organization, 2017: Data quality review: Module 1: framework and metrics, ISBN 9789241512725, https://apps.who.int/iris/handle/10665/259224, last accessed 2021/04/10.

20. World Health Organization, 2020: Overview of the Data Quality Review (DQR) Framework and Methodology, https://cdn.who.int/media/docs/default-source/data-quality-pages/who-dqr-framework-v1-0-overview.pdf?sfvrsn=280bb67_5&sequence=1&isAllowed=y, last accessed 2021/04/10.