Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200139

Data Quality Challenges in a Learning Health System

Michail SARAFIDIS [a], Marilena TAROUSI [a], Athanasios ANASTASIOU [a, 1], Stavros PITOGLOU [b], Efstratios LAMPOUKAS [b], Athanasios SPETSARIAS [b], George MATSOPOULOS [c] and Dimitrios KOUTSOURIS [a]
[a] *Institute of Communication and Computer Systems (ICCS), Athens, Greece*
[b] *Research & Development Dpt., Computer Solutions SA, Athens, Greece*
[c] *School of Electrical & Computer Engineering, National Technical University of Athens, Athens, Greece*

Abstract. This paper discusses the topic of data quality, which concerns the global research and business community and constitutes a challenging task. The data quality prerequisite becomes even more critical when it pertains to critical and sensitive data, such as the healthcare domain data. To begin with, the paper outlines the basic definitions and concepts of data quality and its dimensions. The related research work on data quality assessment is presented and our approach for data quality assurance is introduced. This approach is implemented in our designed cloud platform, called MODELHealth, which is intended for supporting clinical work and administrative decision-making process.

Keywords. data quality, health data, ehr, quality assurance

1. Introduction

Today's world is entirely data-driven. Enormous amounts of data are generated at every point of interaction. Therefore, how this data can be used effectively is a matter of great importance. For this data to be utilizable, they have to meet some requirements and be reliable, up-to-date, and of high quality. Thus, data quality is one of the biggest challenges in the digital world. The high volume and complexity of data collected across multiple sources create an enormous task in terms of data quality management. It is essential to mention that the issues in data quality can directly affect analytics and decision-making.

For these reasons, data quality is recognized as widely considered to be a critical factor, especially in the healthcare sector. This paper outlines the basic definitions and concepts of data quality and proposes a new approach for data quality assurance.

1 Athanasios Anastasiou, Institute of Communication and Computer Systems (ICCS), Zografou Campus, 9 Iroon Polytechniou Str., 15780, Athens, Greece; E-mail: aanastasiou@biomed.ntua.gr.

2. Quality of Health data

2.1. Data Quality Definition and dimensions

Data are real-world objects, with the property of storing, retrieving and elaborating through an algorithmic process and can be communicated between two entities. **Data quality** has a distinct definition in various fields and periods. Multiple definitions of data quality have been proposed, but data is generally considered high quality if **they are fit "for intended uses in operations, decision making, and planning"** [1] or, more generally, **"for the purposes data consumers want to apply them"** [2]. Indeed, the quality of data is critical for improvement process activities, and it can be addressed in various fields inclusive of finance, statistics, computer science, and medicine.

The primary phase in any data quality process is to define the properties of data quality; those properties are defined as data quality dimensions (DQDs). A **Data Quality Dimension** is a term used to define a data quality measure that can describe numerous data elements including attribute, record, table, system or more abstract groupings, such as business unit, company or product range [3]. Researchers have recognized various numbers of dimensions for data quality, but the most important are the six following core dimensions: Accuracy, Completeness, Consistency, Timeliness, Uniqueness and Validity.

Accuracy of the data is the degree to which data is correct, reliable and certified. In other words, data are accurate in case of data values stored in the database are equivalent to real-world values. It is a measure of the correctness of the content of the data.

Completeness of the data concerns the extent to which values appear in a dataset. With regard to a separate datum, only two conditions are potential: The attribute under consideration has a value or not.

Suraj Juddoo et al. suggested that the two common most significant DQDs to consider in the context of big datasets for the health domain are accuracy and completeness [4].

Consistency of the data is the extent to which information is presented in the identical format and compatible with previous data.

Timeliness of data regards only the interval between an alteration of a real-world state and the outcoming modification of the information system state. Timeliness has two constituents: age and volatility, which are a measure of how old the information is and a measure of information instability, respectively.

Uniqueness of data determines the extent to which there are no duplicate values.

Validity of data refers to data that have been collected in conformity with any rules or definitions that apply to that data. This will enable benchmarking between organizations and over time.

According to the "*Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data*" [5], an initiative aimed at providing a solution to the inconsistency of data quality terminology used throughout the relevant scientific literature, the main proposed categories of healthcare data quality are:

Conformance: The data should be compliant as far as format, relational reference, and computational accuracy are concerned. The subcategories of conformance are:

- **Value Conformance**: Values and format of the data should adhere to all given prespecified data architecture constraints.

- **Relational Conformance**: Data should be consistent with the given structural constraints (nullability, referential integrity, etc.).
- **Computational Conformance**: If there are computed values derived from existing data, they should be correct and consistent between computations based on the same specifications (internal or external programs).

Completeness: In accordance with the respective dimension mentioned above, data should be complete, i.e., conforming to the expected frequencies of every data attribute value presence. In this context, completeness is not concerned with the data values, structure or plausibility.

Plausibility: The data values should be believable/truthful when put in context (other variable values, temporal sequences, state transitions etc.). Plausibility subcategories are:

- **Uniqueness Plausibility**: Objects represented by data should not appear multiple times where duplication is not explicitly justified.
- **Atemporal Plausibility**: Observed data values and/or distributions should agree with domain knowledge, trusted external sources or established "gold standards".
- **Temporal Plausibility**: Time-dependent variable values should change as expected, taking under consideration established temporal properties.

2.2. Importance of Data Quality in Electronic Health Record

E-health constitutes a recognized term and is close related to the usage of **Electronic Health Records** (EHRs), so as to achieve adequate communication between healthcare sectors. Specifically, various sources of healthcare data exist, such as the internet, electronic patient records, data analysis tools, communication between health experts and patients, electronic health devices. Data management and analysis tools aim to improve healthcare organizations operation in terms of obtaining useful information and knowledge via conversion of the aforementioned vast resources [6].

It is crucial that the collected data are precise, comprehensive, reliable, comprehensible and available to all stakeholders such as patients, health experts, legal authorities, and government authorities [7]. Therefore, data quality issue constitutes a significant challenge concerning health records and is able to provoke pivotal effects on every health sector operation. A considerable amount of lethal medical errors concerns the result of missing or incomplete data about medication, instructions, and treatment plans [8].

In addition, adequate data quality plays a significant role in the successful implementation of machine learning methods, and it presents one of the significant challenges, especially in the healthcare domain [9], [10]. Poor quality input data are, most of the times, a critical liability in the machine learning process as, in essence, they carry an altered state of the underlying patterns that represent the "truth" of the data, leading to an altered "reality" perceived by the algorithm. It's self-evident that regarding a learning process, if the foundation -the learning material- is critically impaired, one should not have many expectations of the outcome.

The most common source of data used in machine learning processes of medical significance is operational software systems in clinical environments, mainly EHR transactions. This secondary analysis setting, meaning the use of the data for research analytics, which is not their original purpose of collection, has been researched

extensively towards assessing and dealing with the factors that affect data quality [11]–[13]. It has been found that *"EHR data from clinical settings may be inaccurate, incomplete, transformed in ways that undermine their meaning, unrecoverable for research, of unknown provenance, of insufficient granularity, and incompatible with research protocols"* [11].

3. Health Data Quality in MODELHealth

MODELHealth is a cloud platform that facilitates the application of Machine Learning methods, towards processing health data, in order to support clinical work or/and the administrative decision-making process, in all levels of health services provision. It is a "holistic" approach to the subject of developing and exploiting Machine Learning algorithms. It covers the whole lifecycle of the respective process, from querying, homogenizing, anonymizing and enriching raw data, to the final provision of the resulting algorithms via Application Program Interfaces, in order for them to be consumed by any authorized Information System. For projects like MODELHealth, data retrieval is a critical step, as the raw data that are produced by querying a real-world database need to be handled on various levels for a number of reasons before they end up on the central data repository, in order to participate in the training and validation of machine learning algorithms.

To this end, the ETL (Extract, Transform & Load) mechanism of the MODELHealth platform contains a **dedicated Quality Assurance (QA) module**, as well as an Entity Mapping module that handles ontologies and central database homogeneity and an anonymization module that handles patient privacy and regulatory compliance. Both latter modules lie outside the scope of this text.

Regarding the data quality Assessment, the QA module of the ETL system is capable of performing a number of tests on the incoming data, based on a customizable and extendable rule-set that quantifies data quality in the form of test success-failure percentages. There will be rules covering the whole aforementioned spectrum of healthcare data quality: atemporal completeness, value conformance, and atemporal plausibility. These are the factors that are primarily investigated through data quality checks by the organizations which are currently engaged in data quality assessment [14]. Via centrally defined customizable thresholds for each active rule, the administrators can set the quality tolerance of the data retrieval process, thus controlling the minimum quality requirements of the data that actually make it to the central data repository. This is accomplished by "instructing" the decentralized ETL mechanisms to reject data that do not meet the desired criteria.

More specifically, in order to pursue the quality standards, the developed QA module will be based on the research conducted by Nicole G. Weiskopf et al. [15]. The tool's core framework takes into account three different forms of data: complete, correct, and current data. Data quality assessment issues diversify from one another forms of data. Furthermore, the operationalization of the above forms varies depending on the three main types of HER data: patients, variables, and time. As a result, the combination of the aforementioned 3x3 elements produces nine constructs that lead to different methodological recommendations, in other words guideline-based best practices, for EHR data quality assessment.

4. Conclusions

This paper has described a spectrum of issues associated with the quality of health-care data. Because contemporary EHR systems suffer from many shortcomings, the medical data that they produce are also often flawed. Data-quality problems can compromise the value of databases for scientific research, quality assessment, public health, and other purposes. Analytical insights in healthcare should always be probed for data quality problems. Also, the value derived from the use of analytics should dictate the requirements of data quality. This is emphasized because most analytical tools assume that the data are of very high quality. Based on this premise, MODELHealth attempts to apply contemporary data quality checks, towards a systematic approach to data quality, in order to ensure a reliable and viable developed asset for healthcare organizations for the holistic implementation of machine learning processes.

Acknowledgment

This research has been co-funded by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under call Research-Create-Innovate (Project Code: T1EDK-04066).

References

[1] T. C. Redman and D. Ph, Data Driven : Profiting from Your Most Important Business Asset. 2008.
[2] Data International, DAMA-DMBOK : data management body of knowledge, 2nd ed. Technics Publications, 2017.
[3] F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, 'Data quality: A survey of data quality dimensions', 2012 Int. Conf. Inf. Retr. Knowl. Manag., pp. 300–304, Mar. 2012.
[4] S. Juddoo and C. George, 'Discovering Most Important Data Quality Dimensions Using Latent Semantic Analysis', 2018 Int. Conf. Adv. Big Data, Comput. Data Commun. Syst., pp. 1–6, 2018.
[5] M. G. Kahn et al., 'A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data', eGEMs (Generating Evid. Methods to Improv. patient outcomes), vol. 4, no. 1, p. 18, 2016.
[6] Y. Sun, T. Lu, and N. Gu, 'A method of electronic health data quality assessment: Enabling data provenance', Proc. 2017 IEEE 21st Int. Conf. Comput. Support. Coop. Work Des. pp. 233–238, 2017.
[7] M. Botha, A. Botha, and M. Herselman, 'Data quality challenges: A content analysis in the e-health domain', 2014 4th World Congr. Inf. Commun. Technol. WICT 2014, pp. 107–112, 2014.
[8] R. L. Leitheiser, 'Data quality in health care data warehouse environments', Proc. Hawaii Int. Conf. Syst. Sci., vol. 00, no. c, p. 152, 2001.
[9] S. Pitoglou, 'Machine Learning in Healthcare, Introduction and Real World Application Considerations', Int. J. Reliab. Qual. E-Healthcare, vol. 7, no. 2, pp. 27–36, Apr. 2018.
[10] V. N. Gudivada, A. Apon, and J. Ding, 'Data Quality Considerations for Big Data and Machine Learning : Going Beyond Data Cleaning and Transformations', vol. 10, no. 1, pp. 1–20, 2017.
[11] W. R. Hersh et al., 'Caveats for the use of operational electronic health record data in comparative effectiveness research.', Med. Care, vol. 51, no. 8 Suppl 3, pp. S30-7, Aug. 2013.
[12] J. S. Brown, M. Kahn, and S. Toh, 'Data quality assessment for comparative effectiveness research in distributed data networks.', Med. Care, vol. 51, no. 8 Suppl 3, pp. S22-9, Aug. 2013.
[13] A. Rusanov, N. G. Weiskopf, S. Wang, and C. Weng, 'Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research', BMC Med. Inform. Decis. Mak., vol. 14, no. 1, p. 51, Dec. 2014.
[14] T. J. Callahan et al., 'A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks', eGEMs (Generating Evid. Methods to Improv. patient outcomes), vol. 5, no. 1, p. 8, Jun. 2017.
[15] N. G. Weiskopf, S. Bakken, G. Hripcsak, and C. Weng, 'A Data Quality Assessment Guideline for Electronic Health Record Data Reuse', eGEMs (Generating Evid. Methods to Improv. patient outcomes), vol. 5, no. 1, p. 14, Sep. 2017.