Healthcare Transformation with Informatics and Artificial Intelligence
J. Mantas et al. (Eds.)

doi:10.3233/SHTI230414

Data Quality and Data Quantity: Complements or Contradictions?

Jürgen STAUSBERG[a,1] and Sonja HARKENER[a]
[a] *University Duisburg-Essen, Faculty of Medicine, IMIBE, Essen, Germany*
ORCiD ID: Jürgen Stausberg https://orcid.org/0000-0003-0348-5579

Abstract. Although data quality is well defined, the relationship to data quantity remains unclear. Especially the big data approach promises advantages of volume in comparison with small samples in good quality. Aim of this study was to review this issue. Based on the experiences with six registries within a German funding initiative, the definition of data quality provided by the International Organization for Standardization (ISO) was confronted with several aspects of data quantity. The results of a literature search combining both concepts were considered additionally. Data quantity was identified as an umbrella of some inherent characteristics of data like case and data completeness. The same time, quantity could be regarded as a non inherent characteristic of data beyond the ISO standard focusing on the breadth and depth of metadata, i.e. data elements along with their value sets. The FAIR Guiding Principles take into account the latter solely. Surprisingly, the literature agreed in demanding an increase in data quality with volume, turning the big data approach inside out. A usage of data without context - as it could be the case in data mining or machine learning - is neither covered by the concept of data quality nor of data quantity.

Keywords. Data quality, data quantity, metadata, registries

1. Introduction

Data quality is well defined by the International Organization for Standardization (ISO) as "the degree to which a set of inherent characteristics of data fulfils requirements" [1]. Three elements have to be instantiated to operationalize data quality according to ISO 8000, a) characteristics, b) requirements, and c) degree. Dimensions like completeness, correctness, concordance, plausibility and currency have been proposed as characteristics of data (e.g. in [2]). For quality management purposes, characteristics can be specified via quality indicators like "missing values in mandatory data elements" [3]. Requirements are concretizations of use-case-specific needs, e.g. a selection of data elements that is proved by the above mentioned quality indicator. Furthermore, requirements might include a predefined threshold as 5 %, distinguishing between poor and good data quality for the use case at hand [1]. The actual result is provided by calculating the degree in a sample, here the percentage of missing values for the chosen mandatory data elements. Only the adjective "inherent" remains nebulous, briefly

[1] Corresponding Author: Prof. Dr. med. Jürgen Stausberg, Institute for Medical Informatics, Biometry and Epidemiology, Faculty of Medicine, University Duisburg-Essen, Hufelandstrasse 55, 45122 Essen, Germany; e-mail: stausberg@ekmed.de.

introduced as being opposed to "assigned", meaning existing in the data or object[1]

Data have to be regarded as an element of a value chain (cf. figure 1) to be able to define requirements. Data are recorded due to a task within research or health care; data are used for example to assess the efficacy of a new drug in comparison to standard therapy or to detect a contraindication of a prescribed drug based on known conditions of a patient. The result of the value chain could be the approval of a new drug or the withdrawal of a prescription. ISO 8000 does not offer a perspective beyond this value chain, which might occur in case of data mining or machine learning attempts.

The availability of large data samples led to an increasing interest in the quantity of data. For example, national administrative data covering millions of patients could be used for transnational comparisons, registries collect data from whole populations, and the concept of Big Data promises findings without context. Then, why bother about data quality? In our study, we wanted to elaborate the relationship between data quality and data quantity. If possible, we intended to develop an idea how to extend the notion about data quality in ISO 8000 in order to cope issues of data quantity as well.

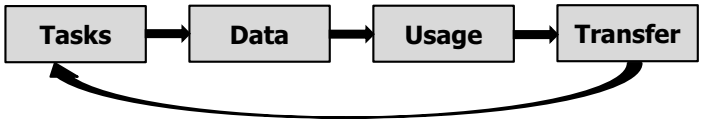

Figure 1. Data within a value chain.

2. Methods

Background of this work is a funding initiative of the German Ministry of Research and Education for six registries. As part of an accompanying project, we were responsible for supporting registry development, implementation and operation (cf. [4] for details). A main concern was related to a cross-registry-benchmarking of data quality using some quality indicators from an available indicator set [3]. Additionally, we performed a scoping review using Medline via www.pubmed.gov. We got 31 hits with the query "(data quantity[Title/Abstract]) AND (data quality[Title/Abstract])" published between 1994 and 2022. The full text was obtained from five publications that seemed relevant for our objectives. Finally, three were considered for this work [5-7].

3. Results

3.1. Data quantity as an inherent characteristic of data

Several quality indicators out of the 51 in [3] are related to some aspects of data quantity. The indicators "recruitment rate" and the rate of "observational units with follow-up" specify case completeness, the indicators "missing values in data elements" and "missing modules" specify data completeness. Both aspects, case and data completeness, are assigned to the dimension completeness of data quality in the literature (e.g. [2]). One can conclude that - in this perspective - data quantity is an inherent characteristic of data quality if the data are part of a value chain (cf. figure 1).

3.2. Data quantity as a not inherent characteristic of data

Within the ISO 8000 framework, the option of data quantity being a not inherent characteristic of data remains. Early attempts to define this concept date back to 1937, "such terms as subtlety, complexity, and intangibility should be recognized as being not inherent characteristics of data." [8]. Unfortunately, the authors did not know an up-to-date definition of "not inherent data quality".

Instead, the metadata collection from the German funding initiative gave an important clue. As an example, the data element "occupation group" is available from one registry only. The other five registries do not record that information. Therefore, their data samples are not fit-for-use if someone is interested in the occupation group from the covered condition. Looking at the ISO 8000, one would not conclude that the data quality is low. The availability of a specific data element is not an inherent characteristic of a data sample. It is rather a consequence of the initial goals of the data sample, or it was forgotten at the time of the data specification. Interestingly, the list of quality indicators includes one measure related to metadata, the "coverage of metadata from investigations" [3]. One can conclude that data quantity represents the volume of metadata, being a non inherent characteristic of data.

3.3. Perspectives of the literature

Two of the three papers defined data quantity as the number of observational units, i.e. case completeness. Msaouel mentioned "large studies" [7], Kolossa/Kopp referred to the "number of subjects" [5]. Additionally, Mayer-Schönberger/Ingelsson picked up the aspect of metadata in saying that "small data samples always miss information – either because it was purposefully not collected or because the random sample just was not large enough to include the salient data points" [6]. Furthermore, they mentioned an exponentially growth of "the absolute number of data points" with the increase in the number of subjects and data elements. Both definitions of data quantity as inherent and non inherent characteristic of data were confirmed by the literature. However, data completeness was not taken into account as an interpretation of quantity.

Msaouel and Kolossa/Kopp concluded a need for better data quality with an increase in volume. Msaouel described the risk of missing the true value with confidence intervals in big data studies naming it the "big data paradox" [7]. One of his recommendations is linked to metadata by quoting for well-defined value sets that cope the heterogeneity of larger samples. Kolossa/Kopp varied the number of subjects in contrast to the number of data points in their simulation [5]. They recommended focusing on data quality by increasing the number of data points instead of the number of subjects. From the point of view of Mayer-Schönberger/Ingelsson, small sample sizes had been always a matter of costs [6]. However, considering a trade-off between data quality and data quantity, they quote a new significance of those who manage data in big data studies.

4. Discussion and conclusions

Data quantity can be regarded as an inherent characteristic of data and consequently as a subtheme to data quality according to ISO 8000. The same time, data quantity can be regarded as a non inherent characteristic of data beyond ISO 8000. With the latter, the perspective of ISO 8000 is supplemented by considering the breadth and depth of metadata. Consequently, one might consider consulting further standards as ISO/IEC

11179-3 “Metadata registries (MDR)”. The quoted gap for data quality in the FAIR Guiding Principles [9] can accordingly be explained by the focus of FAIR on non inherent characteristics of data.

As inherent characteristics of data, data and case completeness complement other characteristics such as plausibility and validity. Data management strategies as the setup of a central monitoring facility will have positive effects on several data quality issues including data quantity. Limited resources might necessitate prioritization, for example to abstain from a case payment to reach a high number of subjects but to perform a source data verification increasing the internal validity. The other way round, missing data can be compensated by statistical procedures to some extent [10].

Struggling for the volume of metadata is a challenge. Forgotten or purposefully abandoned data elements cannot be compensated in case of respective needs in most cases. With this regard, inherent and non inherent characteristics of data might appear as contradictions. In their review, Ramasamy and Chowdhury propose a combination of all characteristics of data [11]. On the one hand, dimensions of data quality are considered as accuracy and integrity. On the other hand, the dimension “metadata” is explicitly introduced under the umbrella term usability, as well as “fitness” under the umbrella term relevance. However, also this review closes with the notion, that “with larger volumes it becomes more important to focus on the quality in order to derive some meaningful insights out of the available data.”

Acknowledgements

The German Federal Ministry of Education and Research partially funded this work under contract 01GY1917B.

References

[1] ISO 8000-2:2020(en) Data quality - Part 2: Vocabulary. https://www.iso.org/obp/ui/#iso:std:iso:8000:-2:ed-4:v1:en (accessed October 11, 2022).
[2] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20:144-51.
[3] Stausberg J, Bauer U, Nasseh D, Pritzkuleit R, Schmidt CO, Schrader T, Nonnemacher M. Indicators of data quality: review and requirements from the perspective of networked medical research. GMS Med Inform Biom Epidemiol. 2019;15:Doc05.
[4] Stausberg J, Harkener S, Semler S. Recent trends in patient registries for health services research. Methods of Information in Medicine. 2021;60 (S01):e1-e8.
[5] Kolossa A, Kopp B. Data quality over data quantity in computational cognitive neuroscience. Neuroimage. 2018;172:775-85.
[6] Mayer-Schönberger V, Ingelsson E. Big Data and medicine: a big deal? J Intern Med. 2018;283:418-29.
[7] Msaouel P. The Big Data Paradox in Clinical Practice. Cancer Invest. 2022; 40: 567-76.
[8] George A. Lundberg and Margaret Lawsing. The Sociography of Some Community Relations. American Sociological Review. 1937;2:318-35.
[9] Stausberg J, Harkener S, Jenetzky E, Jersch P, Martin D, Rupp R, Schönthaler M. FAIR and Quality Assured Data - The Use Case of Trueness. Stud Health Technol Inform. 2022;289:25-8.
[10] Heymans MW, Twisk J. Handling Missing Data in Clinical Research. J Clin Epidemiol. 2022. Online ahead of print. doi: 10.1016/j.jclinepi.2022.08.016.
[11] Ramasamy A, Chowdhury S. Big data quality dimensions: a systematic literature review. Journal of Information Systems and Technology Management. 2020;17:e202017003.