MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230935

FAIR+R: Making Clinical Data Reliable Through Qualitative Metadata

Caroline BÖNISCH[a,1], Dorothea KESZTYÜS[a] and Tibor KESZTYÜS[a]
[a] *Medical Data Integration Center, Department of Medical Informatics, University Medical Center Göttingen, Robert-Koch-Str. 40, 37075 Göttingen, Germany*
ORCiD ID: Caroline Bönisch https://orcid.org/0000-0001-7169-6090, Dorothea Kesztyüs https://orcid.org/0000-0002-2166-846X, Tibor Kesztyüs https://orcid.org/0000-0003-0813-2393

Abstract. Metadata are often the first access to data repositories for researchers within secondary use. Through automatic metadata generation and metadata harvesting the amount of data about data has been growing ever since. In order to make data not only FAIR but also reliable, the aspect of metadata quality has to be considered. But as earlier assessments of metadata of different repositories showed, metadata quality still lacks behind its capability. Providing an extensive literature review the authors conclude nine measures to assess metadata in relation to clinical care repositories, such as Medical Data Integration Centers (MeDICs). Proceeding from these measures the authors propose an addition of the FAIR Guiding Principles by adding a fifth block for Reliability including three principles, that resulted from the measures presented. The results form the basis for the future work of an assessment of metadata, that is stored in a MeDIC.

Keywords. FAIR, metadata, data quality, reliable data

1. Introduction

Since the FAIR Principles were introduced in 2016 [1] the commitment to make data FAIR has increased in different scientific fields [2]. The FAIR principle advise data stewardship and make data Findable, Accessible, Interoperable and Reusable (FAIR), making it particularly applicable in the health area. Since the introduction of the Principles, several initiatives and work groups have formed in order to apply the FAIR Principles in the medical research area. This is necessary because medical data would not be reused for research, although it already exists but is not accessible or findable [3]. However, the FAIR Principles include not only data but also corresponding metadata. Considering the vastly growing collection of data in the field of clinical care, and the establishment of so-called Medical Data Integration Center (MeDIC) at different university hospitals in Germany, the data should not only be findable, accessible, interoperable, and reusable, but also reliable. Only reliable data can form the scientific base of data analysis and provide a potential to validate the results originating from these analysis.

[1] Corresponding Author: Caroline Bönisch, M.Sc., Medical Data Integration Center, Dep. of Medical Informatics, University Medical Center Göttingen Von-Siebold-Straße 3, 37075 Göttingen, Germany; email: caroline.boenisch@med.uni-goettingen.de.

It means in effect that the quality of the data has to be captured and assessed on a scientific premise. The stored data also has to be protected from unplanned changes, being they organizational, structural or content-related. If the data has to be changed, all adjustments must be transparent and stored along the data, comparable to audit trails for electronic medical records [4].

Metadata as accompanying information to the specific data inherit major aspects in providing reliability. In some repositories, data can only be accessed via their metadata and this information is a starting point in secondary use to give researchers a first impression of the data. By adding further details about the quality of the data and metadata being of good quality themselves, the reliability of the information is secured [5].

Metadata consist of intrinsic metadata, for example version number, title, authors, date of creation, and provenance metadata, meaning information about access rights concerning the data or organization hosting the data [6]. Previous literature shows that the quality of data is not fully available within metadata of clinical data [7]. While Ochoa et al. [8] show an overview of different metrics for metadata quality in repositories, they also provide insights of the difference between manual quality evaluation and simple statistical quality measurements and conclude that the quality of metadata should be measured automatically. However, the metrics are short of in the field of multimedia metadata [8].

2. Methods

In the first step, a literature search was executed in order to review already existing evaluation schemes and methods for (meta)data quality.

2.1. Literature Search

Embase via Ovid and PubMed were searched using appropriate search steps and keywords. Table 1 shows the search steps exemplary for the PubMed database.

Table 1. Search Steps of the literature review including count of results for each step in PubMed database.

Number	Search Step	Results
#1	"data accuracy"[MeSH Terms]	3,786
#2	"metadata"[MeSH Terms]	507
#3	"data curation"[MeSH Terms]	816
#4	"quality improvement"[MeSH Terms]	32,640
#5	#1 OR #2 OR #3 OR #4	37,478
#6	("quality"[Title/Abstract] OR "reliable"[Title/Abstract]) AND "metadata"[All Fields]	990
#7	#5 AND #6	136

The search process followed a deductive top-down approach, initially using very abstract search terms, which were then further refined. The PRISMA statement ("Preferred Reporting Items for Systematic Reviews and Meta-Analyses") [9] was used as a conceptual guide of the literature review and the foci of the literature search results were analyzed regarding the following criteria:

- data collected and stored within the clinical environment
- methods or evaluation schemes for estimating data quality

- analysis of (meta)data quality factors
- earlier approaches to make data reliable in other disciplines

Based on the criteria and including the results of the literature search (331 results), title/abstract screening was performed first. The screening resulted in the exclusion of 279 articles and the remainder of 52 articles were then studied completely and finally three were omitted within the full-text review. The publications included were reviewed with a focus on the research criteria of this paper.

2.2. Summarizing Metadata Quality Factors

The results of the literature review revealed different approaches to assess metadata or data quality. As Stausberg et al. [7] stated, the FAIR Principles lack the dimension of quality in (meta)data, therefore the authors additionally examined quality factors for data and adopted them, where possible, to the metadata domain, to provide a complete collection of metadata quality factors for assessment.

Inferred from the metadata quality factors found [10] and presented, the FAIR Principles are then extended with a block for reliable data (RL) and outlined with three principles for this block.

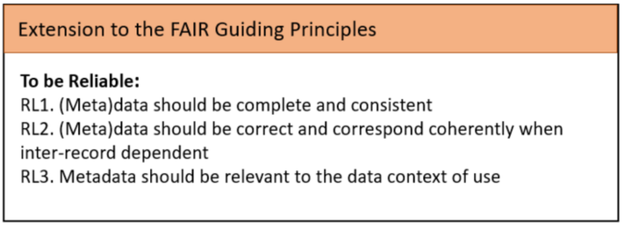
3. Results

Based on the results of the literature review and the related work found, the authors propose the quality measures listed in Table 2. The measures are consolidated from various research manuscripts in different scientific fields and selected concerning the requirements of the data complexity within clinical care.

Table 2. Assessment metrics for metadata in clinical care, based on the results of the literature review.

Measure	Description
Completeness	All mandatory data fields are filled with information
Consistency	Metadata should be conform to existing standards and formats
Correctness	The information describes the metadata in an accurate and distinct way
Correspondence	Metadata that is linked or inter-dependent represents the same information through every instance
Relevance	The metadata corresponds to the requirement/expectations of the user
Semantic Specificity	Average specificity of a semantic concept in metadata information
Timeliness	Currency of the metadata information describing a resource information
Accessibility	The information of the metadata must be physically available and understandable either by human or machine
Reproducibility	Metadata quality scores should be reproducible and not lack clarity in terminology

In conclusion, the authors propose an extension of the FAIR Principles by adding the Principle of Reliability. Figure 1 depicts the three principles, which are added for (meta)data to be Reliable.



Extension to the FAIR Guiding Principles

To be Reliable:
RL1. (Meta)data should be complete and consistent
RL2. (Meta)data should be correct and correspond coherently when inter-record dependent
RL3. Metadata should be relevant to the data context of use

Figure 1. Proposed principle block of Reliable (meta)data.

Using these proposed measures the respective metadata quality can then be calculated. The results of the calculations can subsequently be grouped using the categories Reliable, Reliable with restriction and Not reliable, to provide a straightforward classification of outcomes of a future automatic quality assessment.

4. Discussion

As seen by the literature review, the topic of metadata quality and transferability of qualitative metadata being a key component to reliable data, lacks further research. A highly accepted definition of metadata quality is still somewhat missing as of today, due to the fact that there exists some data quality metrics definition in several scientific areas like bibliography, but little literature results [7], for metadata quality in clinical research could be obtained [11]. Therefore, the literature search had to be extended, to access metrics within data quality, in hopes to apply them to metadata.

Most research regarding metadata, expressed metadata regarding completeness of the data as important quality factor. Nevertheless, it is not the only important quality factor, although maybe the easiest to be measured. Other factors like relevance, consistency and timeliness are also target components of qualitative metadata.

The metadata is on the one hand machine-generated, like date or version, but on the other hand humanly entered via different clinical professionals. Additionally, reusing human-generated data without questioning, when the creator is an expert in the field of research but not an expert in metadata creation, results in discrepancy, as stated by Masor [3].

It should be emphasized that, research of the value and quality of metadata still lags behind the metadata's possibilities. Masor showed that metadata are not being used to their full potential [3]. This is particularly concerning because they are often the first entry point for researchers who want to reuse data from a repository, such as Medical Data Integration Centers (MeDICs) .

As of today, metadata quality assurance is still seen as more of a casualty, and research on this topic is limited. However, as repositories grow, quality issues in metadata gain more visibility and influence the usage of repositories of clinical data.

5. Conclusions

The present article aims to provide an overlook of the existing literature of metadata and data quality concerning clinical care repositories and data integration centers. As the literature on this topic is sparse, further scientific areas were included in order to obtain a complete overview of quality factors of metadata, which are crucial for reliable data.

The quality measures include Completeness, Consistency, Correctness, Correspondence, Relevance, Semantic Specificity, Timeliness, Accessibility, and Reproducibility. Results of these measures can then be classified in reliable, reliable with restriction and not reliable categories, to aid researcher in judging metadata quality.

Based on this aggregation of factors the authors propose an extension of the FAIR Guiding Principles by adding the block of Reliability with additional principles in correspondence to the identified factors.

In accordance with this research, future work will include the automatisation of the metadata quality assessment. This assessment should then be performed on the clinical data collected within the MeDIC of the University Medical Center Göttingen (UMG) and give an overview on the metadata quality of this data collection.

Acknowledgements

Funded by the German Federal Ministry of Education and Research (FAIRRMeDIC, grant number 01ZZ2012).

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar;3(1):160018, doi: 10.1038/sdata.2016.18.

[2] Stall S, Robinson E, Wyborn L, Yarmey L, Parsons M, Lehnert K, Cutcher-Gershenfeld J, Nosek B, Hanson B. Enabling FAIR data across the Earth and space sciences. EOS. 2017 Dec;98, doi: 10.1029/2017EO088425.

[3] Dugas M, Jöckel KH, Friede T, Gefeller O, Kieser M, Marschollek M, Ammenwerth E, Röhrig R, Knaup-Gregori P, Prokosch HU. Memorandum “open metadata”. Methods Inf Med. 2015 Jul;54(04):376-8, doi: 10.3414/ME15-05-0007

[4] Masor JL. Electronic medical records and E-discovery: with New technology come New challenges. Hastings Sci & Tech LJ. 2013;5:245.

[5] Bruland P, Doods J, Storck M, Dugas M. What information does your ehr contain? Automatic generation of a clinical metadata warehouse (CMDW) to support identification and data access within distributed clinical research networks. Stud Health Technol Inform. 2017 Jan;245:313-7, doi: 10.3233/978-1-61499-830-3-313.

[6] Canham S, Ohmann C. A metadata schema for data objects in clinical research. Trials. 2016 Dec;17(1):557, doi: 10.1186/s13063-016-1686-5.

[7] 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 Jan;289:25-8, doi: 10.3233/SHTI210850.

[8] Ochoa X, Duval E. Automatic evaluation of metadata quality in digital repositories. Int J Digit Libr. 2009 Aug;10:67-91, doi: 10.1007/s00799-009-0054-4.

[9] Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009 Jul;6(7):e1000097, doi: 10.1371/journal.pmed.1000097.

[10] Bruce TR, Hillmann DI. The continuum of metadata quality: defining, expressing, exploiting. Metadata in Practice. ALA Editions; 2004.

[11] Shang N, Weng C, Hripcsak G. A conceptual framework for evaluating data suitability for observational studies. J Am Med Inform Assoc. 2018 Mar;25(3):248-58, doi: 10.1093/jamia/ocx095.