Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230121

Local Data Quality Assessments on EHR-Based Real-World Data for Rare Diseases

Kais TAHAR [a,1], Raphael VERBUECHELN [b], Tamara MARTIN [c], Holm GRAESSNER [c] and Dagmar KREFTING [a]
[a] *Institute of Medical Informatics, University Medical Center Göttingen, Germany*
[b] *Medical Data Integration Center, University Hospital Tübingen, Germany*
[c] *Centre for Rare Diseases, University Hospital Tübingen, Germany*

Abstract. The project "Collaboration on Rare Diseases" CORD-MI connects various university hospitals in Germany to collect sufficient harmonized electronic health record (EHR) data for supporting clinical research in the field of rare diseases (RDs). However, the integration and transformation of heterogeneous data into an interoperable standard through Extract-Transform-Load (ETL) processes is a complex task that may influence the data quality (DQ). Local DQ assessments and control processes are needed to ensure and improve the quality of RD data. We therefore aim to investigate the impact of ETL processes on the quality of transformed RD data. Seven DQ indicators for three independent DQ dimensions were evaluated. The resulting reports show the correctness of calculated DQ metrics and detected DQ issues. Our study provides the first comparison results between the DQ of RD data before and after ETL processes. We found that ETL processes are challenging tasks that influence the quality of RD data. We have demonstrated that our methodology is useful and capable of evaluating the quality of real-world data stored in different formats and structures. Our methodology can therefore be used to improve the quality of RD documentation and to support clinical research.

Keywords. Data quality, rare disease, healthcare standards, ETL, HL7 FHIR

1. Introduction

The research project "Collaboration on Rare Diseases" (CORD-MI) [1] of the German Medical Informatics Initiative (MII) [2] connects multiple university hospitals to support clinical research with electronic health record (EHR) data in the field of rare diseases (RD). In Europe, RD are defined as diseases that affect less than 5 in 10,000 people [3], therefore multi-site data sharing is required to reach a sufficient number of cases. However, the required integration and transformation of heterogeneous data sources through Extract-Transform-Load (ETL) processes into an interoperable format is a complex and challenging task [4]. Such ETL processes raise concerns about data quality (DQ) issues such as completeness, implausibility and semantic integrity of the final data sets [5]. To ensure an appropriate evidence level of scientific outcomes derived from these data, sufficient DQ is necessary [6]. The source of a potential DQ issue, i.e. the primary documentation or the ETL processes, can be determined by evaluating the DQ before and after ETL processes, as presented in this manuscript.

1 Corresponding Author: Kais Tahar, Georg-August-University, University Medical Center Göttingen, Institute of Medical Informatics, 37075 Gottingen, Germany; E-mail: kais.tahar@med.uni-goettingen.de.

Various data quality frameworks have been proposed in the literature [5]–[8]. However, useful DQ assessments on RD data usually depend on specific user and domain requirements [6]. The aim of the present work is to evaluate the quality of RD data before and after ETL processes using a methodology that takes user and domain specific requirements into consideration as proposed in [6]. In this context, we investigate how the quality of real-world data on RDs can be assessed automatically, which methods and tools can be used to compare the quality of real-world data stored in different data formats and structures, which impacts have ETL processes on the quality of RD data, and report the learned lessons.

2. Methods

We describe the employed data sets and DQ metrics before we present the implemented methods in more detail. Since the terms in related works about DQ are often ambiguous, we use in this paper the terminology presented in [6].

2.1. Data sets and data sources

The base population used for this study encompasses all inpatient cases in 2020 of a university hospital, stored in the patient administration system (PAS). PAS is the central subsystem in the hospital information systems (HIS) responsible for patient admission, patient discharge and medical billing. In effect, all clinical subsystems send the captured patient data to the PAS - in our case the SAP IS-H system [9] is used. From the base population, all cases that were coded with an ICD-10-GM code [10] covered in a reference list are included into the study. This reference list is also available on GitHub repository [11]. It includes 143 diagnoses that can be used for coding RDs. The extracted data items are specified in the MII core data set (MII-CDS) [12]. The MII-CDS defines the semantics of required data items and provides the basis for enabling standardized data exchange as well as harmonized DQ assessments across the CORD-MI network [6].

The selected data sets for this study are stored in Fast Healthcare Interoperability Resources (FHIR) [13] and comma-separated values (CSV) formats. Both data sets capture information about the basic modules of the MII-CDS namely Person, Treatment Case, and Diagnosis [12]. Exemplary data sets in CSV and FHIR are provided in [6, 11]. The first data set (SAP data) used for this study was exported directly from SAP IS-H in CSV format. The second data set (FHIR data) was created using an ETL pipeline that extracts clinical data from SAP IS-H according to the German §21 Hospital Remuneration Act (KHEntgG) [14] and transforms these data into FHIR standard. The resulting FHIR resources follow the MII-CDS as specified in the FHIR implementation guide of CORD-MI [15]. These standardized data are stored in a central FHIR server that provides multiple types of FHIR resources such as Patient, Encounter, and Condition.

2.2. Data quality concept and assessment methods

In this study, three DQ dimensions were considered, namely completeness, plausibility, and uniqueness - dimensions defined as most relevant together with domain experts in CORD-MI [6]. Five DQ parameters (parameters relevant for DQ but not indicating DQ) and seven DQ indicators are derived from these dimensions as shown in tables 1 and 2.

RD-specific coding with so-called Orphacodes (OCs) is necessary to avoid any ambiguity in RD documentation. Specific metrics are therefore used to assess the quality of RD data. For definitions of used DQ metrics please refer to [6].

The DQ library (dqLib) [16] has been used to develop specific reporting scripts for DQ assessments. This R package provides methods that enable users to select desired dimensions, indicators, and parameters as well as to define specific DQ reports [6]. Using this software framework, a specific DQ tool was implemented to import the employed data sets and configure required DQ reports [11]. To compare the quality of RD data before and after ETL processes, we applied the developed tools on the two data sets. The generated DQ report comprises two Excel spreadsheets for each data set. The first sheet illustrates the calculated DQ metrics as shown in tables 1 and 2, while the second sheet reports the detected DQ issues. All reports are checked manually for contradictions. If no contradiction can be found the reports are considered as correct.

3. Results

Table 1. DQ parameters displayed in the generated reports for 2020. The resulting RD cases are unambiguously identified by OCs or tracer ICD-10-GM codes. All rel. frequencies are normalized to 100.000 cases.

Data Set	Inpatient Cases	Analyzed Inpatient Cases	RD Cases rel. Frequency	Orpha Cases rel. Frequency	Tracer Cases rel. Frequency
SAP	79810	1415	649	538	241
FHIR	79810	1417	221	0	221

Table 2. DQ indicators displayed in the generated reports for 2020.

DQ Dimension	DQ Indicator	Abr.	SAP Data	FHIR Data
Completeness (co)	Item Completeness Rate	dqi_co_icr	85,71%	78,57%
	Value Completeness Rate	dqi_co_vcr	99,48%	95,64%
	Orphacoding Completeness Rate	dqi_co_ocr	53.73%	0
Plausibility (pl)	Orphacoding Plausibility Rate	dqi_pl_opr	93,88%	NA
	Range Plausibility Rate	dqi_pl_rpr	100%	100%
Uniqueness (un)	RD Case Unambiguity Rate	dqi_un_cur	94,02%	93,18%
	RD Case Dissimilarity Rate	dqi_un_cdr	50%	100%

Tables 1 and 2 present the results of DQ assessments performed on the employed data sets. Discrepancies between the DQ results obtained before and after ETL can be observed in most parameters and indicators. The FHIR data has no Orpha-coded cases and as a consequence less RD cases than the SAP data as illustrated in table 1. Table 2 shows that both indicators for item and value completeness are notably higher in SAP data than that in FHIR data. Furthermore, the unambiguity rate of RD cases is slightly higher in SAP data than in the FHIR data. However, the dissimilarity indicator achieved better results on FHIR data and reached its maximal level after the ETL processes.

4. Discussion

Table 1 shows a discrepancy between the parameters obtained before and after the ETL processes. Both data sets have the same origin (SAP IS-H) and show the same number of inpatient cases. However, the data in the PAS is not static but underlies corrections even on historical data. This may explain the little discrepancy found in Analyzed

Inpatient Cases and Tracer Cases rel. Frequency. It is a bit counterintuitive that the analyzed inpatient cases are higher in FHIR data, but the tracer cases are lower. However, as the German remuneration act allows for retrospective clarifications, recording in the PAS may still occur for previous years. Furthermore, the FHIR data does not contain any data items for capturing OCs, because the §21 Act did not support this standard in 2020 [14]. Only diagnoses with an unambiguous ICD-10-GM code (so-called tracer) allow the identification of RD cases. The legislature has responded to the necessity of better RD documentation and from April 2023 on, adding OCs according to Alpha-ID-SE terminology [17] will become mandatory for all RD documentation in §21 data.

Table 2 shows that the completeness indicators performed better on SAP data than on FHIR data. The main reason for that are FHIR mapping errors that have been identified using the DQ reports and will be therefore removed in the next update. Another reason is the lack of required OCs in the FHIR data as mentioned above. In contrast, the dissimilarity indicator performed better on FHIR data. This indicator shows that the SAP data contains duplicated RD cases. Using ETL the data were cleansed from duplications and transformed into a standardized format. Therefore, the dissimilarity indicator reached its maximal level after ETL processes. In addition, table 2 shows that although the FHIR data do not contain any OCs, our methods were able to compute the RD Case Unambiguity Rate only based on available ICD-10-GM codes. The ETL did not introduce any issues regarding the range plausibility.

The execution of DQ assessments runs without errors. There are no contradictions in the generated DQ reports. The study results have therefore indicated the correctness of calculated DQ metrics and detected DQ issues. The implemented metrics cover independent aspects of DQ [6]. Our study has shown that the developed methodology is capable of detecting potential DQ issues such as missings, implausibility or ambiguity of RD diagnoses and that it can be used for reporting on the quality of real-world data stored in heterogeneous data sources. The DQ reports helped us to compare the DQ before and after ETL processes and to find the causes of detected DQ violations. The results were validated independently by domain experts. The used DQ dimensions and metrics fit well the specified requirements, with certain limitations as described in [6]. In future works we will apply our methods on data stored in distributed data sources across multiple hospitals to compare the quality of RD data recorded in different HISs.

5. Conclusion

The resulting DQ reports have shown the correctness of calculated DQ metrics and detected DQ issues. Our work is the first study to investigate the impact of ETL processes on the quality of RD data to our knowledge. We showed that our methodology is able to identify DQ issues by comparing the DQ before and after ETL processes. We found discrepancies between the DQ results obtained before and after ETL processes, some of them based on errors in the transformation step. Such complex and challenging processes can decrease the quality of RD data. On the other hand, our study has shown that the extracted data were cleansed from duplications and transformed into an interoperable standard using ETL processes. This enhances the reuse of RD data for clinical research. We have demonstrated the usefulness and portability of developed tools by applying our methodology to real-world data stored in different data formats and structures. Our methodology can therefore be used to improve the quality of RD documentation and to support clinical research.

Ethical Approval: This study was performed in line with the principles of the Declaration of Helsinki. The analysis of retrospective, pseudonymized data collected for patients with rare disease diagnosis at University Hospital Tübingen in 2020, using the technical infrastructure of the MI-initiative, was approved by the ethics committee of University Hospital Tübingen (reference number 514/2020BO2).

Author Contributions: Conceptualization, methodology, and writing-original draft: KT and DK; requirement analysis and definition of indicators: KT, DK, TM, and HG; software for data quality assessments: KT; use case execution and software for data curation: KT and RV; writing-review and editing: KT, RV, DK, TM, and HG.

Acknowledgement: This study was done within the CORD-MI project funded by the German Federal Ministry of Education and Research (BMBF), funding numbers 01ZZ1911R and 01ZZ1911O.

References

[1] 'Use Case CORD-MI | Medizininformatik-Initiative'. https://www.medizininformatik-initiative.de/de/CORD (accessed Jun. 10, 2023).

[2] Semler SC, Wissing F, Heyder R. German medical informatics initiative. Methods of information in medicine. 2018 May;57(S 01):e50-6. doi: 10.3414/ME18-03-0003.

[3] Martin T et al., 'Problems of finding rare diseases in the documentation of German hospitals', Sep. 2021, p. DocAbstr. 59. doi: 10.3205/21gmds117.

[4] Tahar K, et al. Integrating heterogeneous data sources for cross-institutional data sharing: requirements elicitation and management in SMITH. InMEDINFO 2019: Health and Wellbeing e-Networks for All 2019 (pp. 1785-1786). IOS Press., doi: 10.3233/SHTI190647.

[5] Spengler H, et al. Improving data quality in medical research: a monitoring architecture for clinical and translational data warehouses. In2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS) 2020 Jul 28 (pp. 415-420). IEEE. doi: 10.1109/CBMS49503.2020.00085.

[6] Tahar K. et al., 'Rare Diseases in Hospital Information Systems – An Interoperable Methodology for Distributed Data Quality Assessments', *Methods Inf Med*, vol. 0, no. AAM, doi: 10.1055/a-2006-1018.

[7] Ramasamy A. and Chowdhury S. 'Big Data Quality Dimensions: A Systematic Literature Review', J. Inf. Technol. Manag., vol. 17, no. 0, Art. no. 0, May 2020, doi: 10.4301/S1807-1775202017003.

[8] Schmidt CO, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Medical Research Methodology. 2021 Dec;21(1):1-5. doi: 10.1186/s12874-021-01252-7.

[9] 'SAP Software Solutions | Business Applications and Technology'. https://www.sap.com/index.html (accessed Dec. 30, 2022).

[10] 'BfArM - ICD-10-GM'. https://www.bfarm.de/EN/Code-systems/Classifications/ICD/ICD-10-GM/_node.html (accessed Jan. 10, 2023).

[11] Tahar K., Data Quality Assessment on Rare Diseases Data – A set of Metrics and Tools for the 33rd Medical Informatics Europe Conference 2023. https://doi.org/21.11101/0000-0007-FB71-F

[12] 'The MII core data set'. https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set (accessed Jan. 10, 2023).

[13] Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. InProceedings of the 26th IEEE international symposium on computer-based medical systems 2013 Jun 20 (pp. 326-331). IEEE. doi: 10.1109/CBMS.2013.6627810.

[14] 'Datenübermittlung an das InEK.' https://www.dkgev.de/themen/digitalisierung-daten/elektronische-datenuebermittlung/datenuebermittlung-an-das-inek/ (accessed Dec. 29, 2022).

[15] 'CORD-MI ImplementationGuide'. https://simplifier.net/guide/medicalinformaticsinitiative-cord-implementationguide (accessed Dec. 23, 2022).

[16] Tahar T, Data Quality Library (dqLib): R package for data quality assessment and reporting. 2022. https://doi.org/21.11101/0000-0007-F6DE-A (accessed Dec. 29, 2022).

[17] 'BfArM Alpha-ID-SE'. https://www.bfarm.de/EN/Code-systems/Terminologies/Alpha-ID-SE/_node.html (accessed Jan. 10, 2023).