Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210904

Towards Harmonized Data Quality in the Medical Informatics Initiative – Current State and Future Directions

Matthias LÖBE[a,1], Gaetan KAMDJE-WABO[b], Adriana Carina SINZA[c], Helmut SPENGLER[d], Marcus STROBEL[e] and Erik TUTE[f]

[a] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Germany*
[b] *Center for Preventive Medicine and Digital Health, Department of Biomedical Informatics, Medical Faculty Mannheim of the University of Heidelberg, Germany*
[c] *Data Integration Center, University Hospital Aachen, Germany*
[d] *Institute for Artificial Intelligence and Informatics in Medicine, Technical University of Munich, Germany*
[e] *Data Integration Center, University Hospital Leipzig, Germany*
[f] *Peter L. Reichertz Institute for Medical Informatics (PLRI), Hannover Medical School, Germany*

Abstract. Health data from hospital information systems are valuable sources for medical research but have known issues in terms of data quality. In a nationwide data integration project in Germany, health care data from all participating university hospitals are being pooled and refined in local centers. As there is currently no overarching agreement on how to deal with errors and implausibilities, meetings were held to discuss the current status and the need to develop consensual measures at the organizational and technical levels. This paper analyzes the discovered similarities and differences. The result shows that although data quality checks are carried out at all sites, there is a lack of both centrally coordinated data quality indicators and a formalization of plausibility rules as well as a repository for automatic querying of the rules, for example in ETL processes.

Keywords. Data Quality, Electronic health record, Medical Informatics Initiative

1. Introduction

Electronic health record (EHR) data from health care information systems have particular, well-recognized weaknesses in the area of data quality due to background of their collection (treatment, not completeness, as the primary purpose, limited human resources for documentation in hospitals, complex real-life processes with divergences between medical treatment and technical recording) [1]. As a result, some clinical researchers, as well as biometricians, have reservations about using EHR data for

[1] Corresponding Author, Matthias Löbe, Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany; E-mail: matthias.loebe@imise.uni-leipzig.de.

research and do not see them as equivalent to prospectively collected data backed by data management processes as in clinical trials.

In the German medical informatics initiative (MII) [2], data integration centers (DIC) have been established since 2018 at currently 29 university hospitals, which aim to make healthcare data available for research. The DICs are each assigned to one of four consortia[2], whereby each consortium is based on different technical information architectures. To ensure a comparable data corpus in all DICs, a central core data set was agreed upon, which contains modules from different areas (demographics, encounter, diagnoses, procedures, laboratory data, medications, etc.). The core data set is specified in HL7 FHIR format.

In the last years, the DICs have focused primarily on connecting sources to assemble a comprehensive data pool. Aspects of data quality have not been a focus of the work. However, for the broad and partially automated sharing of the data envisaged for the future, measures have to be developed that ensure the data quality required for each use case. This was exemplified in the so-called MII demonstrator study [3].

2. Methods

Within a joint working group, a series of questions was developed and sent to designated representatives of the consortia. These questions queried aspects of six complexes: 1) the organizational structures set up locally, 2) the technical infrastructure, 3) the handling of errors in the process of extracting, transforming, and loading from the primary information systems, 4) the curation processes in the research database, 5) the plan for developing further functionalities, and 6) the envisaged need for central conventions and consensual data quality rules. Responses were presented and discussed in a workshop.

3. Results

The topic ‘data quality’ was recognized in its importance by all consortia and addressed in a surprisingly comparable depth. There are local working groups that deal with conceptual, implementation-related and organizational aspects. These groups are also networked with other groups within the consortia (e.g., those responsible for data extractions from primary systems), but the data quality assurance procedures outlined are not yet universally used in routine operations.

However, the implementation status of data quality measures varies considerably *within* individual sites of a consortium. For the most part, pilot sites exist that also have lead responsibility for developing the concepts and tools. It was noticeable that beside the long scientific history, existing data quality approaches were only used to a limited extent. New technical developments are used in the majority of cases. Available information systems, e.g., from the area of clinical trials or cohorts, which address data quality problems technically [4], e.g., through dedicated query management or data curation boards, are not used. This may be due in part to the fact that the size of the project required the use of many staff from fields other than medical informatics. As the ETL pipelines mature and are more widely used, a deeper understanding of the structure and limitations of the local data bodies is now currently emerging.

[2] https://www.medizininformatik-initiative.de/en/consortia/data-integration-centres

Finally, it should be noted that the researchers' view of "errors" must also be questioned. Not every case of complaint are real errors, because in reality, the hospital data are recorded at a very detailed level, which has to serve different purposes of use. This includes provisional values, cancellations, error corrections, recoding, and similar operations that are not present in the smoothed view that is usually presented to researchers.

Differences between the consortia naturally concern the technical implementation. Since the underlying grant program was competitively bid, the information architecture is the same within consortium sites, but different between the consortia. The technical architecture of the ETL pipeline in DIFUTURE integrates comprehensive event logging through which data quality is viewed in a structured and detailed manner [5]. An audit service allows flexibly configurable quality analyses, which are executed on a SQL-based data mart but are independent of concrete schemas.

In HiGHmed, there is a special focus on data governance. Comprehensive organizational structures have been established including data stewards for modeling clinical concepts of a domain, responsible parties for each source system, and a data reviewing board for overall, regular analysis of data sets. The in-house development openCQA [6] makes commonly governed compilations (e.g. for reports or dashboards) of various data quality indicators applicable on HiGHmed's technical architecture.

MIRACUM stores important data quality indicators and rules into a metadata repository (MDR) [7]. The self-developed software DQAstats generates detailed, cross-site standardized error reports in PDF format on a quarterly basis to ensure, monitor and document the measures taken [8]. These reports are published anonymously in the consortium for comparison and self-assessment.

The SMITH consortium has designed a five-stage data quality assurance concept, which, starting with manual tests based on a coordinated catalog of data quality indicators and at defined intervals, continuing with automated procedures (in development), also plans for the use of central terminology services, the connection of a metadata repository and natural language processing of free text annotations.

4. Discussion and Future Work

In summary, promising approaches have been developed that now need to be rolled out across the range of consortia sites, put into operation, and feedback incorporated. Nevertheless, objective evidence of the qualitative suitability of the extracted data for the variety of potential research projects is still lacking. This will require the involvement of a broader community of domain experts from other areas of biomedical research such as biometrics and epidemiology, as they have already addressed a variety of similar problems and developed strategies to solve them. Another need is seen in the training and further qualification of staff, which will lead to a more effective involvement of specific and harmonized Data Stewards abilities across the different consortia. Furthermore, it is considered to construct a plausible clinical question to query real data and to test known typical error constellations on the shared data in a cross-consortium 'projectathon'. Those involved in the workshop agreed to use the framework of Kahn et. al. [9] as a taxonomy of error types.

In the longer term, common solutions are to be developed in three sub-areas. First, this concerns the development of a system of harmonized data quality indicators and rules for their operationalization. This approach should separate the conceptual specification from the syntax for execution in order to be able to support different target platforms. It seems worthwhile to define common data quality indicators for certain common data elements of the MII core data set (e.g., LOINC Top 300 most common laboratory values) and also to define metrics or thresholds for quality assurance in data quality assessments.

Secondly, a repository as a storage and access location for the harmonized rules would be desirable. In all consortia, a Metadata Repository is already in productive use or at least announced. In principle, these systems could also be used to manage the data quality indicators and rules as outlined by MIRACUM, if suitable programming interfaces for mutual access or syndication become available. Further workarounds, such as the making use of some GIT-based repositories could also for this purpose be investigated, so that a distributed and location-wide access is ensured towards a continuous MDR-based assessment of data quality indicators.

The third issue relates to the availability of reference data sets for data validation. Many advanced data quality issues cannot be identified by looking at locally available data alone. Comparison to a gold standard is needed. This gold standard can vary widely in nature and scope; as an example, a comparison of the frequency of distribution of certain numbers of cases or answer categories between the local site and a larger (and thus more statistically robust) number of cases would be useful in identifying systemic biases. With this in mind, combined, i.e., aggregated reference data from all MII sites would be beneficial.

Acknowledgement

The work is supported by the corresponding BMBF/MII grants (01ZZ1803A, 01ZZ1801E, 01ZZ1803B, 01ZZ1804A, 01ZZ1803D, 01ZZ1802C, 01ZZ1910A).

References

[1] Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res 2021;23(3):e22219. doi: 10.2196/22219.

[2] Semler S, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018;57(S 01):e50-e56. doi: 10.3414/ME18-03-0003.

[3] Kamdje-Wabo G, Gradinger T, Löbe M, Lodahl R, Seuchter SA, Sax U et al. Towards Structured Data Quality Assessment in the German Medical Informatics Initiative: Initial Approach in the MII Demonstrator Study. Stud Health Technol Inform 2019;264:1508–9. doi: 10.3233/SHTI190508.

[4] Löbe M, Meineke F, Winter A. Scenarios for Using OpenClinica in Academic Clinical Trials. Stud Health Technol Inform 2019;258:211–5. doi: 10.3233/978-1-61499-959-1-211.

[5] Spengler H, Lang C, Mahapatra T, Gatz I, Kuhn KA, Prasser F. Enabling Agile Clinical and Translational Data Warehousing: Platform Development and Evaluation. JMIR Med Inform 2020;8(7):e15918

[6] Tute E, Scheffner I, Marschollek M. A method for interoperable knowledge-based data quality assessment. BMC Med Inform Decis Mak 2021;21(1):93. doi: 10.1186/s12911-021-01458-1.

[7] Kadioglu D, Breil B, Knell C, Lablans M, Mate S, Schlue D et al. Samply.MDR - A Metadata Repository and Its Application in Various Research Networks. Stud Health Technol Inform 2018;253:50–4.

[8] Kapsner LA, Kampf MO, Seuchter SA, Kamdje-Wabo G, Gradinger T, Ganslandt T et al. Moving Towards an EHR Data Quality Framework: The MIRACUM Approach. Stud Health Technol Inform 2019;267:247–53. doi: 10.3233/SHTI190834.

[9] Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC) 2016;4(1):1244. doi: 10.13063/2327-9214.1244.