R. Röhrig et al. (Eds.)

doi:10.3233/SHTI190834

Moving Towards an EHR Data Quality Framework: The MIRACUM Approach

Lorenz A. KAPSNER[a,1,2], Marvin O. KAMPF[a,2], Susanne A. SEUCHTER[a], Gaetan KAMDJE-WABO[c], Tobias GRADINGER[c], Thomas GANSLANDT[c], Sebastian MATE[a], Julian GRUENDNER[b], Detlef KRASKA[a] and Hans-Ulrich PROKOSCH[a,b]

[a]*Center of Medical Information and Communication Technology, University Hospital Erlangen, Germany;*
[b]*Chair of Medical Informatics, Friedrich-Alexander University Erlangen-Nürnberg, Germany;*
[c]*Department of Biomedical Informatics of the Heinrich-Lanz-Center, Mannheim University Medicine, Ruprecht-Karls-University Heidelberg, Germany*

Abstract. Introduction: Data quality (DQ) is an important prerequisite for secondary use of electronic health record (EHR) data in clinical research, particularly with regards to progressing towards a learning health system, one of the MIRACUM consortium's goals. Following the successful integration of the i2b2 research data repository in MIRACUM, we present a standardized and generic DQ framework. State of the art: Already established DQ evaluation methods do not cover all of MIRACUM's requirements. Concept: A data quality analysis plan was developed to assess common data quality dimensions for demographic-, condition-, procedure- and department-related variables of MIRACUM's research data repository. Implementation: A data quality analysis (DQA) tool was developed using R scripts packaged in a Docker image with all the necessary dependencies and R libraries for easy distribution. It integrates with the i2b2 data repository at each MIRACUM site, executes an analysis on the data and generates a DQ report. Lessons learned: Our DQA tool brings the analysis to the data and thus meets the MIRACUM data protection requirements. It evaluates established DQ dimensions of data repositories in a standardized and easily distributable way. This analysis allowed us to reveal and revise inconsistencies in earlier versions of the ETL jobs. The framework is portable, easy to deploy across different sites and even further adaptable to other database schemes. Conclusion: The presented framework provides the first step towards a unified, standardized and harmonized EHR DQ assessment in MIRACUM. DQ issues can now be systematically identified by individual hospitals to subsequently implement site- or consortium-wide feedback loops to increase data quality.

Keywords. Data analysis, data quality, electronic health record, clinical research

[1] Corresponding Author: Dr. med. Lorenz Kapsner, Center of Medical Information and Communication Technology, University Hospital Erlangen, Krankenhausstraße 12, 91054 Erlangen, Germany; E-mail: lorenz.kapsner@uk-erlangen.de.

[2] These authors contributed equally to this work.

1. Introduction

1.1. Background

Data quality is an important prerequisite for the secondary use of electronic health record (EHR) data [1,2], especially for establishing a learning health system, one of the *MIRACUM* consortium's goals [3]. MIRACUM is a collaboration of 10 university hospitals and industry partners in Germany aiming to overcome the challenges of digitalization and future research in medicine, as part of the *German Medical Informatics Initiative* [4]. Within MIRACUM's initial concept phase, Haverkamp et al. demonstrated the deployment of an i2b2 [5] data repository, the execution of a distributed research query and the subsequent analysis of the data across eight MIRACUM partner sites [6]. Another pilot study implemented and evaluated the OMOP [7] common data model (CDM) and the data analysis tools of the *Observational Health Data Sciences and Informatics* program (OHDSI) [8] across MIRACUM [9].

Following the successful integration of the i2b2 research data repository in MIRACUM, we present a framework to evaluate its data quality in a standardized and versatile manner.

1.2. Requirements

The MIRACUM data quality analysis (DQA) framework should assess common dimensions of EHR data quality like *conformance, completeness and plausibility* [1,10]. These concepts are also part of the data quality guidelines [11] of the TMF, a German platform for collaborative research in medicine. According to Nonnemacher, Nasseh and Stausberg [11], they can be summarized as the following quality indicators: missing values for data elements (TMF-ID 1013), data elements with unknown values (TMF-ID 1016), invalid values for qualitative data elements (TMF-ID 1021), invalid values for quantitative data elements (TMF-ID 1024) and consistency (TMF-1003). Furthermore, principles such as those described by Huebner et al. in their publication about initial data analysis in the context of studies with primary-data collection [12] should be taken into account. Portability, distribution across all MIRACUM sites and easy integration into a hospital's research IT infrastructure must be considered. The DQA framework should be easy to use and the results must be presented in a clearly structured way and reported in at least two different granularities: detailed site-specific results to enable a deeper analysis of data quality irregularities and a summarized report with aggregated statistics, which can be shared with others. To enable broad access and continuous development, it should be provided under an open source license.

2. State of the art

Lee et al. published "a framework for data quality assessment in clinical research datasets" [13]. This article however focuses on the identification of DQ data elements for heart failure and the implementation as a software tool is planned as future work. OHDSI provides *Achilles* [14], an R package that allows one to perform data quality and descriptive analyses on OMOP databases. Bialke et al. developed an R package to support data quality assurance in epidemiological research [15]. Unfortunately, these tools are either bound to particular database schemes or not generic enough to be easily configured to address the same data quality criteria and complex plausibility checks across multiple research repositories (e.g. i2b2 and OMOP) with the same code basis. In order for the tool to be extendable in the future, as the content of our MIRACUM research repositories grows, such a generic approach is indispensable. This allows an easy-to-use DQA tool to be established for all MIRACUM sites. For the proof-of-concept analysis performed during MIRACUM's conceptual phase, data quality has only been verified via random sampling checks at each respective partner site, mostly by means of self-written SQL queries and not based on a systematic consortium-wide data quality analysis plan. Thus, within the MIRACUM consortium, a standardized approach for assessing the data quality of the research data repositories was still required.

3. Concept

In MIRACUM, the already implemented i2b2 repository currently (02/2019) contains pseudonymized billing data in a standardized format. These data include 21 variables related to medical data of inpatient hospital stays. We assessed basic data quality criteria of the i2b2 data repository in one of our previous works [16]. Starting from there, we developed a data analysis plan that covers the variables of the billing data set, including demographic-, condition-, procedure- and department-related variables. Conformance of the data was evaluated by presenting the summary statistics of the different variables in a report. These results can subsequently be compared to the constraints given in the official data dictionary [17], which are also included into the report. Completeness was determined by counting missing values of each variable. To assess plausibility, we checked for incorrect multiple occurrence of data values. These checks were established based on common domain expert knowledge of the participating hospital's data managers and physicians, e.g. *every encounter ID may only be associated with one principal diagnosis* or *every encounter ID may only be associated with one distinct patient ID*. Additionally, plausibility was tested for by investigating relationships between data elements and checking respective values against defined constraints. Two MIRACUM sites in cooperation systematically defined eight relations, as well as the corresponding constraints (see Table 1 and Kamdje-Wabo et al. [18]). We decided to pursue a two-level data quality analysis: first, validating the contents of the final i2b2 research data repository on its own and second, examining the validity of the extract-transform-load (ETL) jobs by comparing the contents of the i2b2 repository with the raw data (CSV files exported from the billing system).

Table 1. Excerpt of the defined relations between data elements and constraints to estimate plausibility.

Plausibility Relation	Constraint
A diagnosis from ICD-10-GM chapter XV (pregnancy, childbirth and puerperium) is only permitted for female patients.	Gender must be female.
Malignant neoplasms of the female genital organs (ICD-10-GM C51-C58) are only permitted as hospital diagnosis in female patients.	Gender must be female.
Malignant neoplasms of the male genital organs (ICD-10-GM C60-C63) are only permitted as hospital diagnosis in male patients.	Gender must be male.
"Inpatient delivery" is only permitted as a reason for admission for female patients.	Gender must be female.
Age in years at first admission must not be negative.	Age in years must be >= 0.

4. Implementation

The DQA tool was implemented as an R script. To develop a generic and versatile tool, we extended this script with custom analysis functions, settings- and layout files. The functions were designed to be applicable both to the i2b2 data repository (target data) and the CSV files (source data). To make the DQA tool transferable to other database schemes, e.g. OMOP, we moved the i2b2 SQL queries to an external file, which is referenced in the script. All necessary components are deployed via a centralized MIRACUM collaboration platform together with a Docker [19] image that contains all system libraries and packages of the open source statistical programming language R [20] that are required to execute the analyses and to generate the data quality report. As graphical user interface (GUI), the open source software RStudio Server [21] was included into the image, to customize e.g. location-specific details of the report or to manually launch the generation of the data quality analysis. The final report consists of a comprehensive document in which the results of the analyses of the source and target systems are compared in a predefined structure. The DQA tool generates log files and stores them on the host system. These can be downloaded and are well suited for deeper manual investigations of potential data irregularities by e.g. data managers that are familiar with the site-specific data. To enable sharing with others, summarized results and aggregated statistics are exported in tabular format. The DQA tool integrates easily with the i2b2 repositories used to conduct the federated analyses published by Haverkamp et al. [6] (see Figure 1).

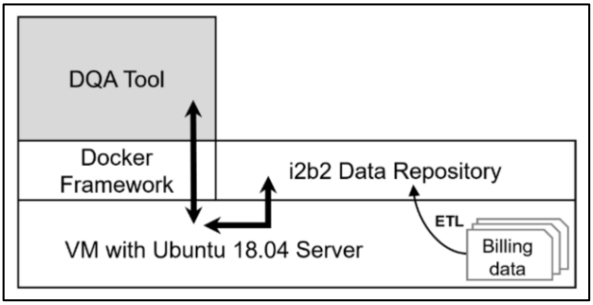

Figure 1: Schematic architecture of the MIRACUM data quality analysis (DQA) tool. The Docker container is additionally installed to a preexisting i2b2 research data repository. The DQA tool can access i2b2's database via the host VM. Legend: VM = virtual machine, ETL = extract-transform-load. Colors: white boxes = host operating system; gray box = Docker container; thick arrows = internal network communication; thin arrow = ETL job.

5. Lessons learned

Our DQA tool assesses commonly used data quality dimensions of MIRACUM's i2b2 data repository in a standardized way. Our approach brings the DQA tool to the data and not vice versa and thus meets the MIRACUM data protection requirements. It is suitable for distribution across the MIRACUM sites and its integration into any hospital's research IT infrastructure is straightforward. Portability to other database schemes is ensured by using generic configuration files. Using the tool is familiar to researchers due to the use of the common RStudio Server GUI. Based on our experience, the deployment of the DQA tool in the above-described approach facilitates a consortium-wide integration. According to our collaboration platform, all ten MIRACUM partners successfully integrated the tool within 20 days following its release. The tool has already been utilized by all ten MIRACUM partners to fulfill their site-specific milestones of creating their first-year data quality reports.

In conclusion, the presented framework meets all requirements stated in section 1.2. Further, the DQA tool will be published under an open source license in the future. While developing the DQA tool, we continuously tested the data quality of the i2b2 repository at the Erlangen University Hospital and revealed inconsistencies in earlier versions of the ETL jobs, which were subsequently analyzed and revised. One example was a logical error in an SQL statement that caused the number of distinct patients in the table containing demographic information to differ from the number of patients in the table containing clinical observations by erroneously skipping patients during the loading process. The emended ETL jobs were subsequently included in a new distribution of the MIRACUM i2b2 data repository.

We are planning to extend the DQA tool to the OMOP data scheme and to integrate descriptive statistical visualizations. Furthermore, more advanced and complex plausibility checks addressing e.g. relationships between procedures and diagnoses or diagnoses and age will be developed and integrated. In the course of the further extension of the core data set from billing data to the successive integration of other EHR data elements, the functionality of the DQA tool will be regularly adapted. We also aim to automate more parts of the tool, e.g. conformance checks. This could be supported by MIRACUM's metadata repository [22] in the future, e.g. by generically applying data quality analyses based on a variable's metadata.

6. Conclusion

We presented a federated DQA framework, which provides the basis for moving towards a unified, standardized and harmonized EHR data quality assessment across MIRACUM. Its generic design makes it adaptable to infrastructures of future partners and other consortia. The MIRACUM partners have already successfully created a first data quality report as proposed by the MIRACUM project plan and are now able to systematically and regularly determine and explore site-specific irregularities and to implement local or consortium-wide feedback loops to increase data quality, e.g. by improving data collection-, ETL- or programming procedures. Understanding the causes of irregularities in MIRACUM's data quality enables us to establish generally accepted criteria for high data quality within our consortium. We are currently in the process of collecting formal feedback from all MIRACUM sites and finalizing a comprehensive evaluation plan for the DQA tool which will then be applied in the upcoming release. Further, we are collecting suggestions for data-element specific thresholds from each partner site, regarding e.g. missing values, to achieve a MIRACUM-specific consensus and based on this define additional data quality checks. To get an overview of the current state of the data quality of MIRACUM's research data repositories, we plan to collect and analyze each site's aggregated results of the

DQA report. This will enable us to continuously improve data quality and to support distributed research analyses with data quality aspects in the future.

Conflict of Interest

The authors state that they have no conflict of interests.

Acknowledgement

This work was funded in part by the German Federal Ministry of Education and Research (BMBF) within the Medical Informatics Initiative (MIRACUM Consortium) under the Funding Numbers FKZ: 01ZZ1801A, 01ZZ1801E. The present work was performed in (partial) fulfillment of the requirements for obtaining the degree “Dr. rer. biol. hum.” for the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (MOK).

References

[1] N.G. Weiskopf, C. Weng, and N.G. Weiskopf, Methods and dimensions of electronic health record data quality assessment : enabling reuse for clinical research, (2013) 144–151. doi:10.1136/amiajnl-2011-000681.

[2] N.G. Weiskopf, G. Hripcsak, S. Swaminathan, and C. Weng, Defining and measuring completeness of electronic health records for secondary use, *J. Biomed. Inform.* (2013). doi:10.1016/j.jbi.2013.06.010.

[3] H.-U. Prokosch, T. Acker, J. Bernarding, H. Binder, M. Boeker, M. Boerries, P. Daumke, T. Ganslandt, J. Hesser, G. Höning, M. Neumaier, K. Marquardt, H. Renz, H.-J. Rothkötter, C. Schade-Brittinger, P. Schmücker, J. Schüttler, M. Sedlmayr, H. Serve, K. Sohrabi, and H. Storf, MIRACUM: Medical Informatics in Research and Care in University Medicine, *Methods Inf. Med.* **57** (2018) e82–e91. doi:10.3414/ME17-02-0025.

[4] S.C. Semler, F. Wissing, and R. Heyder, German Medical Informatics Initiative., *Methods Inf. Med.* **57** (2018) e50–e56. doi:10.3414/ME18-03-0003.

[5] S.N. Murphy, M. Mendis, K. Hackett, R. Kuttan, W. Pan, and LC, Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside, *Symp. A Q. J. Mod. Foreign Lit.* (2007).

[6] C. Haverkamp, T. Ganslandt, P. Horki, M. Boeker, A. Dörfler, S. Schwab, J. Berkefeld, W. Pfeilschifter, W.D. Niesen, K. Egger, M. Kaps, M.A. Brockmann, E. Neumaier-Probst, K. Szabo, M. Skalej, S. Bien, C. Best, H.-U. Prokosch, and H. Urbach, Regional Differences in Thrombectomy Rates: Secondary use of Billing Codes in the MIRACUM (Medical Informatics for Research and Care in University Medicine) Consortium, *Clin. Neuroradiol.* **28** (2018) 225–234. doi:10.1007/s00062-017-0656-y.

[7] J.M. Overhage, P.B. Ryan, C.G. Reich, A.G. Hartzema, and P.E. Stang, Validation of a common data model for active safety surveillance research, *J. Am. Med. Informatics Assoc.* (2012). doi:10.1136/amiajnl-2011-000376.

[8] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Y.-C. Li, P.E. Stang, D. Madigan, and P.B. Ryan, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers., *Stud. Health Technol. Inform.* (2015).

[9] C. Maier, L. Lang, H. Storf, P. Vormstein, R. Bieber, J. Bernarding, T. Herrmann, C. Haverkamp, P. Horki, J. Laufer, F. Berger, G. Höning, H.W. Fritsch, J. Schüttler, T. Ganslandt, H.-U. Prokosch, and M. Sedlmayr, Towards Implementation of OMOP in a German University Hospital Consortium, *Appl. Clin. Inform.* **9** (2018) 54–61. doi:10.1055/s-0037-1617452.

[10] M.G. Kahn, T.J. Callahan, J. Barnard, A.E. Bauck, J. Brown, B.N. Davidson, H. Estiri, C. Goerg, E. Holve, S.G. Johnson, S.-T. Liaw, M. Hamilton-Lopez, D. Meeker, T.C. Ong, P. Ryan, N. Shang, N.G. Weiskopf, C. Weng, M.N. Zozus, and L. Schilling, 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)*. (2016). doi:10.13063/2327-9214.1244.

[11] M. Nonnemacher, D. Nasseh, and J. Stausberg, Datenqualität in der medizinischen Forschung. Leitlinie zum adaptiven Management von Datenqualität in Kohortenstudien und Registern, Schriftenr, Medizinisch Wissenschaftliche Verlagsgesellschaft, Berlin, 2014.

[12] M. Hübner, S. le Cessie, C. Schmidt, and W. Vach, A Contemporary Conceptual Framework for Initial Data Analysis, *Obs. Stud.* **4** (2018) 171–192. doi:10.4049/jimmunol.1102689 ET - 2012/03/07.

[13] K. Lee, N. Weiskopf, and J. Pathak, A Framework for Data Quality Assessment in Clinical Research Datasets., *AMIA ... Annu. Symp. Proceedings. AMIA Symp.* (2017).

[14] P. Ryan, M. Schuemie, V. Huser, C. Knoll, A. Londhe, and T. Abdul-Basser, Achilles: Creates Descriptive Statistics Summary for an Entire OMOP CDM Instance, (2018).

[15] M. Bialke, H. Rau, T. Schwaneberg, R. Walk, T. Bahls, and W. Hoffmann, mosaicQA - A General Approach to Facilitate Basic Data Quality Assurance for Epidemiological Research, *Methods Inf. Med.* (2017). doi:10.3414/me16-01-0123.

[16] M.O. Kampf, D. Kraska, and H. Prokosch, An Analysis of Erlangen University Hospital ' s Billing Data on Utility- Based De- Identification, in: ICT Heal. Sci. Res., 2019: pp. 70–74. doi:10.3233/978-1-61499-959-1-70.

[17] Deutsche Krankenhausgesellschaft, Daten nach § 21 KHEntgG – Version 2011 für das Datenjahr 2010, (2011) 1–36. http://www.dkgev.de/media/file/8729.V-21-KHEntgG_A-2011_2010-12-01_k.pdf (accessed May 20, 2019).

[18] G. Kamdje-Wabo, T. Gradinger, M. Löbe, R. Lodahl, S. Seuchter, and U. Sax, Towards Structured Data Quality Assessment in the German Medical Informatics Initiative: initial approach in the MII Demonstrator Study, (2019).

[19] D. Merkel, Docker: Lightweight Linux Containers for Consistent Development and Deployment, *Linux J.* (2014). doi:10.1097/01.NND.0000320699.47006.a3.

[20] R Core Team, R: A Language and Environment for Statistical Computing, (2018). https://www.r-project.org.

[21] RStudio Team, RStudio: Integrated Development Environment for R, (2015). http://www.rstudio.com/.

[22] D. Kadioglu, B. Breil, C. Knell, M. Lablans, S. Mate, D. Schlue, H. Serve, H. Storf, F. Ückert, T. Wagner, P. Weingardt, and H.U. Prokosch, Samply.MDR - A Metadata Repository and Its Application in Various Research Networks, *Stud. Health Technol. Inform.* (2018).