dHealth 2025
M. Baumgartner et al. (Eds.)

doi:10.3233/SHTI250161

Assuring End-to-End Data Quality for Analytics on FHIR

Jasmin ZIEGLER[a,b,c1], Clara FISCHER[b,e], Paul-Christian VOLKMER[b,d], Marcel ERPENBECK[a], Jonathan M. MANG[a], Thomas GANSLANDT[c], Hans-Ulrich PROKOSCH[b,c] and Christian GULDEN[b,c]

[a] *Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen, Germany*
[b] *Bavarian Cancer Research Center (BZKF)*
[c] *Friedrich-Alexander-Universität Erlangen-Nürnberg, Medical Informatics, Erlangen, Germany*
[d] *Comprehensive Cancer Center Mainfranken, Würzburg, Germany*
[e] *Medical Data Integration Center (MEDIZUKR), University Hospital Regensburg, Regensburg Germany*

Abstract. Background: The accumulation of Real-World Data (RWD) from Electronic Health Records (EHRs) and registries offers substantial potential for generating Real-World Evidence (RWE). However, the ability to generate robust evidence from real-world data hinges on its quality. This is especially critical when heterogeneous data is first transformed into standardized, research-ready data models. Objective: This study presents an approach for assessing data completeness through a pipeline for extracting and transforming oncological RWD. Methods: We introduce a technical solution that enables the assessment of data completeness across three data transformation stages, beginning with the initial data source and extending through Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) to CSV. Results: Using Trino, a distributed SQL engine, we evaluate data completeness at the three transformation stages by comparing cancer diagnosis counts. The modular pipeline design, compatible with various data sources, allows for error detection in ETL processes. Conclusion: Future work will expand the system to address additional data quality dimensions, such as correctness and plausibility, improving the overall robustness of data analytics in federated environments.

Keywords. data quality, observational study, Health Level Seven® FHIR®, electronic health records

1. Introduction

Electronic Health Records (EHRs) and data from registries, as systematic collections of information on specific patient populations and conditions, have resulted in an unprecedented accumulation of Real-World Data (RWD) within the healthcare sector. RWD offers significant potential for deriving retrospective insights into real-world patient populations, ultimately aiming to generate Real-World Evidence (RWE). However, RWD frequently originates from a variety of systems initially collected for non-research purposes, which can pose significant challenges to maintaining data

[1] Corresponding Author: Jasmin Ziegler, Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen, Germany, jasmin.ziegler@uk-erlangen.de

quality. Drawing reliable conclusions from causal inference and generating credible RWE depends on having high-quality and comprehensive RWD [1].

The Bavarian Cancer Research Center (BZKF) has united six university hospitals to advance cancer detection, prevention, diagnosis, and treatment. In addition to conducting multicenter clinical trials, the university hospitals have initiated the Oncology Real World Data Platform (ORWDP), a federated observational research network [2]. Generating RWE from RWD in such a federated environment adds complexity due to varied systems and data. To address the challenge of harmonizing data across multiple institutions, the standardized data model Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) has proven to be an effective approach for achieving data uniformity throughout Germany [3]. In particular, this also allows for integrating data from multiple data sources within the same institution. In this context, we developed a data processing pipeline that transforms oncological RWD to FHIR and ultimately generates tabular datasets tailored to specific research questions for use in federated analyses [4]. This paper describes a proof-of-concept implementation for evaluating the data completeness throughout all transformation stages in this pipeline.

2. Methods

The lower part of figure 1 illustrates a simplified version of the pipeline we developed in our previous work [4] which transforms oncological real-world data into an interoperable data model and prepares the harmonized data for federated analysis. The pipeline consists of the following transformation steps:

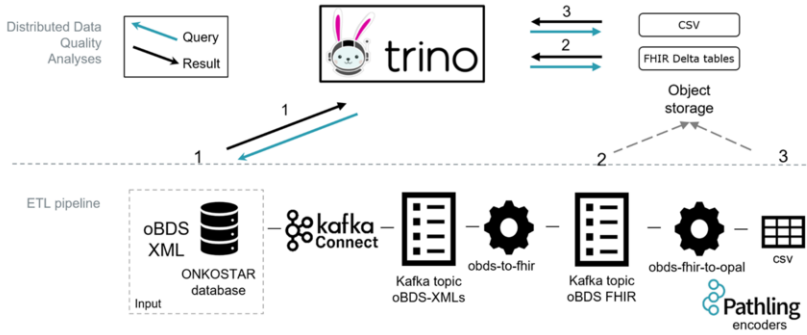

Figure 1. Technical Setup for evaluating data completeness from data source to data sink.

1. Data ingestion: the pipeline ingests standardized XML data (oncological basic dataset, oBDS) from the ONKOSTAR tumor documentation system, which transmits oBDS to cancer registries as required by German law. Apache Kafka Connect streams the data from the ONKOSTAR database.
2. Transformation into a harmonized and interoperable data model: an extract-transform-load (ETL) process (obds-to-fhir [5]) converts the oBDS XML reports into HL7 FHIR, a standard for healthcare data interoperability.
3. Data conversion for analysis: the Pathling library FHIR encoders [6] in the obds-fhir-to-opal service convert the FHIR resources into a CSV file, preparing them for analysis.

To assess the data completeness across the transformation steps, we first limit our queries to records of patients diagnosed with cancer at University Hospital Erlangen, documented in ONKOSTAR and reported to the Bavarian Cancer Registry. We defined an exemplary query to compare the record counts:
Query 1: counts of total cancer diagnoses stratified by year (2018 - 2023)
Query 2: counts of specific cancer diagnoses stratified by ICD10 code (2022)

We use Trino [7], a distributed SQL engine capable of querying data from multiple sources, to compare counts across oBDS reports in the ONKOSTAR database, the FHIR resources, and the final CSV file. We developed SQL scripts for each of the three stages to retrieve the counts for both queries.

Figure 1 further outlines the extended pipeline setup, comparing data completeness at three points. Trino connects directly to the ONKOSTAR database (1). Additionally, we load the FHIR resources encoded as Delta Lake tables (2) and the CSV output (3) into object storage. Using Trino, we connect to all three data sources, run the SQL queries across them and retrieve results as a single table. We calculate the absolute and relative differences between the systems [8].

3. Results

Table 1 compares counts of total cancer diagnoses between aggregated ONKOSTAR, FHIR, and CSV datasets from 2018 to 2023 (query 1). The diagnosis counts between ONKOSTAR and FHIR show minor relative differences, ranging from 0.1 % to 0.63 %. The CSV diagnosis counts align exactly with FHIR for all years. The last digit of the absolute numbers have been replaced by an "x" to maintain data confidentiality, as the relative values sufficiently highlight the differences between the systems.

Table 1. Results for query 1: Comparison of diagnosis counts between ONKOSTAR, FHIR, and CSV data sources (2018-2023)

Year of Diagnosis	ONKOSTAR Diagnosis Count (1)	FHIR Diagnosis Count (2)	Relative Difference between (1) and (2) [%]	CSV Diagnosis Count (3)	Relative Difference between (2) and (3) [%]
2018	583x	583x	0.10	583x	0
2019	535x	536x	0.11	536x	0
2020	509x	509x	0.12	509x	0
2021	500x	501x	0.30	501x	0
2022	488x	491x	0.63	491x	0
2023	415x	414x	0.10	414x	0

Table 2 presents the diagnosis data for 2022, categorized by ICD-10 code (query 2). We selected 2022 as a representative example since it is the most recent year with likely complete data, whereas 2023 and 2024 may still be subject to retrospective documentation adjustments. It highlights relative differences between ONKOSTAR and FHIR for the six most frequently diagnosed entities, ranging from 0 % to 1.50 % (mean relative deviation: 0.60 %) and shows exact alignment between FHIR and CSV counts. For the years 2020, 2021 and 2023, we found similar results (mean relative deviations

2020: 0.11 %, 2021: 0.56 %, 2023: 0.27 %). A demo setup of all components is available online [9].

Table 2. Results for query 2: Counts of cancer diagnoses, categorized by ICD-10 codes, for the six most frequently diagnosed entities during the one-year period of 2022

ICD10 code	ONKOSTAR Diagnosis Count (1)	FHIR Diagnosis Count (2)	Relative Difference between (1) and (2) [%]	CSV Diagnosis Count (3)	Relative Difference between (2) and (3) [%]
C50	57x	57x	0.35	57x	0
C61	40x	40x	0.25	40x	0
C43	36x	36x	0	36x	0
C44	26x	26x	1.50	26x	0
C34	24x	24x	0	24x	0
D06	20x	19x	1.50	19x	0

4. Discussion

Our setup allows for data completeness checks across multiple data sources by leveraging SQL as a common query language across data sources and transformation steps. Despite the complexity through the involvement of various tools, the modular design makes the pipeline adaptable and possibly suitable as a reference implementation for similar projects.

A key aspect of the ETL validation process is having a deep understanding of the source database structures. This, however, can be time-consuming and resource-intensive, especially since most commercial products do not consider their database schemas a public interface. Additionally, there is a strong incentive to avoid fully recreating ETL jobs in SQL for validation purposes, as this could introduce further complexity and potential errors. Instead, it is necessary to acknowledge that some discrepancies may occur due to the intrinsic differences between the source and transformed data. For example, incomplete data in the source system may fail the FHIR validation process, preventing the creation of FHIR resources and leading to a lower number of FHIR diagnosis counts. Additionally, the investigated site migrated to the ONKOSTAR tumor documentation system in 2020/2021, with previously documented data transferred to the new system. Both of these factors could contribute to discrepancies in diagnosis counts between the two transformation stages. Therefore, while a 100% completeness threshold might be ideal, achieving such a standard could be challenging in a real-world observational scenario. Given the complexities introduced by the data migration and potential discrepancies in diagnosis counts, increasing the failure threshold might represent a more realistic and achievable benchmark in terms of data completeness [10].

The data completeness reports presented in this work allow us to quickly identify inconsistencies and pinpoint incomplete data that may have been lost during the transformation process. They can serve as valuable tools for enhancing the development of ETL jobs by providing insights into data transformation processes and identifying areas for improvement. By modifying the SQL queries, we can generate a diverse range of reports tailored to specific analytical research questions, facilitating a deeper understanding of data quality and completeness.

Our practical experience has shown that discrepancies tend to occur more often between the source database and FHIR rather than between FHIR and downstream systems. These insights reinforce the importance of thoroughly understanding and validating the ETL process at each stage to ensure data consistency and reliability.

While our current focus is on ensuring data completeness, future work will involve extending our checks to other important dimensions of data quality such as correctness, concordance, and plausibility [11, 12]. The flexibility of SQL enables the implementation and execution of these more complex queries designed to address specific quality assurance requirements.This approach aligns with the recommendation for a systematic and comprehensive data quality assessment framework for EHR data, as outlined by Kahn et al. [13]. Van der Lei's work further highlights the risks associated with the misuse of data when repurposed for activities beyond its original intent. In such cases, ensuring the integrity of the data is critical [14]. The proposed setup offers a solid foundation for further data quality measures like implementing dashboards, or incorporating additional data quality tools like Great Expectations [15] or the MIRACUM DQA tool [16].

5. Conclusion

Transformations can be a source of data quality issues, especially between formats as dissimilar as real-world data from relational databases and standardized and harmonized FHIR resources. We have shown how a distributed query engine can be used to address this, by providing a single access point to compare source systems with transformation results regarding completeness. This approach allows streamlining the validation process, ensuring that discrepancies are quickly identified and addressed within ETL processes.

Declarations

Conflict of Interest: The authors declare that they have no conflict of interest.
Funding: This project has received funding from the Bavarian Cancer Research Center (BZKF) and support from the BZKF lighthouse AI & Bioinformatics.
The present work was performed in (partial) fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum." from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (JZ).

References

[1] Penberthy LT, Rivera DR, Lund JL, Bruno MA, Meyer A-M (2021) An overview of real-world data sources for oncology and considerations for research. CA Cancer J Clin 72:287–300. https://doi.org/10.3322/caac.21714

[2] Ziegler J, Gruendner J, Rosenau L, Erpenbeck M, Prokosch H-U, Deppenwiese N (2023) Towards a Bavarian Oncology Real World Data Research Platform. Stud Health Technol Inform 307:78–85. https://doi.org/10.3233/SHTI230696

[3] Semler S, Wissing F, Heyder R (2018) German Medical Informatics Initiative. Methods Inf Med 57:50–56. https://doi.org/10.3414/ME18-03-0003

[4] Ziegler J, Erpenbeck M, Fuchs T, Saibold A, Volkmer P-C, Schmidt G, Eicher J, Pallaoro P, De Souza Falguera R, Aubele F, Hagedorn M, Vansovich E, Raffler J, Ringshandl S, Kerscher A, Maurer J, Kühnel B, Schenkirsch G, Kampf M, Kapsner LA, Ghanbarian H, Spengler H, Soto-Rey I, Albashiti F, Hellwig D, Ertl M, Fette G, Kraska D, Boeker M, Prokosch H-U, Gulden C (2024) Bridging Data Silos in Oncology with Modular Software for Federated Analysis on FHIR: A Multisite Implementation Study. JMIR Prepr. https://doi.org/10.2196/preprints.65681

[5] bzkf/obds-to-fhir. https://github.com/bzkf/obds-to-fhir. Accessed 8 Jan 2025

[6] Grimes J, Szul P, Metke-Jimenez A, Lawley M, Loi K (2022) Pathling: analytics on FHIR. J Biomed Semant 13:23. https://doi.org/10.1186/s13326-022-00277-1

[7] Trino | Distributed SQL query engine for big data. https://trino.io/. Accessed 15 Aug 2024

[8] AbuHalimeh A (2022) Improving Data Quality in Clinical Research Informatics Tools. Front Big Data 5:871897. https://doi.org/10.3389/fdata.2022.871897

[9] BZKF paper-e2e-data-quality. https://github.com/bzkf/paper-e2e-data-quality/tree/master. Accessed 20 Jan 2025

[10] Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR (2021) Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc JAMIA 28:2251–2257. https://doi.org/10.1093/jamia/ocab132

[11] Weiskopf NG, Weng C (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 20:144–151. https://doi.org/10.1136/amiajnl-2011-000681

[12] Wang Z, Talburt JR, Wu N, Dagtas S, Zozus MN (2020) A Rule-Based Data Quality Assessment System for Electronic Health Record Data. Appl Clin Inform 11:622–634. https://doi.org/10.1055/s-0040-1715567

[13] Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw S-T, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L (2016) A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs 4:1244. https://doi.org/10.13063/2327-9214.1244

[14] Lei J van der (1991) Use and Abuse of Computer-Stored Medical Records. Methods Inf Med 30:79–80. https://doi.org/10.1055/s-0038-1634831

[15] Great Expectations: have confidence in your data, no matter what • Great Expectations. https://greatexpectations.io/. Accessed 9 Sep 2024

[16] Kapsner LA, Mang JM, Mate S, Seuchter SA, Vengadeswaran A, Bathelt F, Deppenwiese N, Kadioglu D, Kraska D, Prokosch H-U (2021) Linking a Consortium-Wide Data Quality Assessment Tool with the MIRACUM Metadata Repository. Appl Clin Inform 12:826–835. https://doi.org/10.1055/s-0041-1733847