doi:10.3233/SHTI240658

Challenges and Lessons Learned in Mapping HL7 v2 Data to openEHR: Insights from UKSH Medical Data Integration Center

Michael ANYWAR [a,b,1], Mário MACEDO [a,b], Santiago PAZMINO [a,b], Tobias BRONSCH [a,b], Benjamin KINAST [a,b], Ann-Kristin KOCK-SCHOPPENHAUER [b,c] and Björn SCHREIWEIS [a,b]
[a] Institute for Medical Informatics and Statistics, Kiel University and University Hospital Schleswig-Holstein, Germany
[b] Medical Data Integration Center, University Hospital Schleswig-Holstein, Kiel/Lübeck, Germany
[c] IT Center for Clinical Research, Universität zu Lübeck, Lübeck, Germany
Michael Anywar: https://orcid.org/0000-0001-8028-803X, Mário Macedo: https://orcid.org/0009-0001-5309-2194, Santiago Pazmino: https://orcid.org/0009-0009-7065-0221, Tobias Bronsch: https://orcid.org/0000-0002-3622-0625, Benjamin Kinast: https://orcid.org/0000-0003-2554-4381, Ann-Kristin Kock-Schoppenhauer: https://orcid.org/0000-0001-6685-4429, Björn Schreiweis: https://orcid.org/0000-0002-1748-1563

Abstract. This paper explores the challenges and lessons learned during the mapping of HL7 v2 messages structured using custom schema to openEHR for the Medical Data Integration Center (MeDIC) of the University Hospital, Schleswig-Holstein (UKSH). Missing timestamps in observations, missing units of measurement, inconsistencies in decimal separators and unexpected datatypes were identified as critical inconsistencies in this process. These anomalies highlight the difficulty of automating the transformation of HL7 v2 data to any standard, particularly openEHR, using off-the-shelf tools. Addressing these anomalies is crucial for enhancing data interoperability, supporting evidence-based research, and optimizing clinical decision-making. Implementing proper data quality measures and governance will unlock the potential of integrated clinical data, empowering clinicians and researchers and fostering a robust healthcare ecosystem.

Keywords. Interoperability, Data quality control, Healthcare Data Mapping, ETL, openEHR, HL7 v2, Data integration, Health Information Exchange

1. Introduction

In today's healthcare landscape, the seamless exchange of patient data is essential for informed clinical decisions and improved patient care [1]. This exchange is facilitated

[1] Corresponding Author: Michael Anywar, University Hospital Schleswig-Holstein, Hörn Campus, Kaistraße 101, 24114 Kiel, Germany; E-mail: Michael.anywar@uksh.de.

by widely adopted health informatics data standards such as Fast Healthcare Interoperability Resources (FHIR), OMOP CDM [2] and openEHR [3]. However, the complex task of mapping data from legacy standards such as HL7v2 or between these standards poses challenges, particularly with non-standardized or incomplete clinical datasets.

At the University Hospital Schleswig-Holstein (UKSH) the Medical Data Integration Center (MeDIC) aims to consolidate clinical data from various sources such as hospital, laboratory and radiology information systems, as well as picture archiving and communication systems. The MeDIC serves as a research clinical data repository, supporting data-driven studies, epidemiological research, and clinical decision-making [4].

UKSH has been using custom HL7 v2 message structure, hence this work delves into the data quality challenges encountered while mapping both legacy and contemporary HL7 v2 data to openEHR, specificall HL7 v2 ORU Laboratory results and HL7 v2 Bar messages fro Diagnosis and Procedures. By addressing these anomalies, the study aims to provide valuable insights for enhancing interoperability, and data integration. The lessons learned are expected to improve the accuracy of mapping healthcare data between standards, thereby fostering seamless interoperability.

2. Methods

The process of mapping data to openEHR not only involved the transformation process but also included identifying and analyzing data inconsistencies within HL7 v2 data during the mapping [5]. Talend, a data integration platform played a crucial role in orchestrating the data transformations from HL7 v2 to openEHR standard. In Figure 1, UKSH's MEDIC data mapping orchestration is depicted, highlighting Talend's dual focus on mapping to openEHR and handling data inconsistencies.

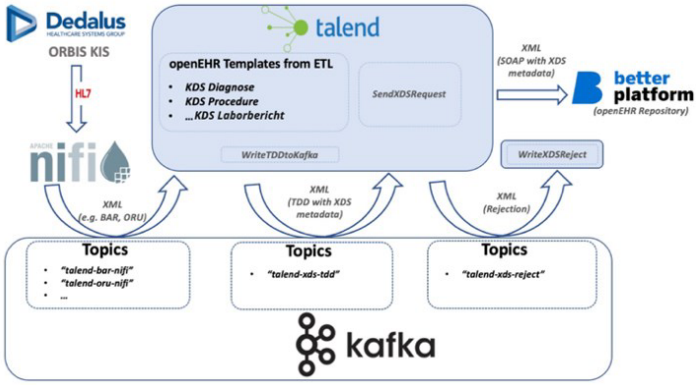

Figure 1. UKSH MeDIC HL7 v2 - openEHR Mapping Orchestration.

A comparative analysis between data stored in the openEHR repository and data indexed on MeDIC's Elasticsearch platform revealed misalignments, prompting further investigation. Subsequently, 1000 candidate HL7 v2 ORU messages and 1000 HL7 v2 BAR messages were selected for reprocessing through Talend jobs, leading to the discovery of several data quality anomalies. The failure of certain messages to be processed in subsequent iterations of the Talend jobs allowed for the identification of

discrepancies, such as missing elements and non-conformities in the UKSH custom HL7 v2 messages. Issues related to observation dates, measurement units, and data types were identified and addressed both within the Talend job iterations and across the organization to ensure conformity in the destination openEHR repository.

3. Results

The data quality analysis on the transformation of HL7 v2 data to openEHR via Talend jobs unveiled significant inconsistencies and challenges. Table 1 summarizes the identified issues and the corresponding actions taken to address them.

Table 1. Summary of Data Quality Issues and Actions Taken

Issue	Description	Action Taken
Missing Observation timestamps	Some HL7 v2 laboratory messages lack timestamps, posing challenges in determining when these observations occurred.	Implemented a systematic process to trace back observations to their source and correct timestamps, ensuring data completeness and accuracy.
Missing Measurement Units	Measurement units for were absent in some laboratory observation data.	Developed procedures to register measurements lacking units and update them within the openEHR repository alongside observations.
Inconsistent Decimal Separators	Inconsistent use of decimal separators (e.g., commas vs. periods) in numerical observations.	Enhanced the mapping pipeline to standardize numerical observations according to German/EU conventions, ensuring consistency in data processing.
Datatype Mismatch	Text-based entries in numeric fields complicated the data mapping process.	Identified and converted text-based entries with numerical value ranges into text-based observation results to align with openEHR repository expectations.
Duplicate Data	Multiple HL7 v2 messages received during an event's lifecycle causing data redundancy.	Implemented message ID reconciliation and update mechanisms in the openEHR repository to manage duplicate data, ensuring data integrity and reliability.

4. Discussion

The result of this work highlights the paramount importance of adhering to health informatics standards and implementing robust data standardization practices in healthcare systems. Even when utilizing custom data formats, as demonstrated by the UKSH with the use of custom HL7 v2 message formats, ensuring data interoperability and bolstering clinical data quality remains imperative. Aligned with the digital health DQ-DO framework [6], which identifies consistency as a pivotal dimension of data quality, our findings accentuate the critical need to prioritize both consistency and completeness in digital health data. These dimensions play a pivotal role, impacting various facets of data quality outcomes, including clinical efficacy, clinician trust, research validity, operational efficiency, and organizational effectiveness. Addressing these dimensions is crucial as they impact various digital health data quality outcomes,

including clinical, clinician [7], research-related, business process, and organizational outcomes [8].

Central to our findings is the recognition of barriers in achieving seamless mapping of HL7 messages to openEHR using off-the-shelf mapping/transformation tools. The dynamic nature of healthcare data in HL7 introduces complexities, requiring constant tool modifications to address anomalies effectively. Healthcare organizations must invest in robust data governance mechanisms, implement rigorous data validation processes, and prioritize data standardization efforts to mitigate data quality issues and support effective healthcare delivery and decision-making.

Lessons Learned

The experience of dealing with missing observation timestamps highlights the importance of a multi-faceted approach in addressing challenges rooted in source data. Proactive reporting and preventive measures enhance data quality and emphasize collaboration across the data pipeline [6].

The absence of measurement units taught valuable lessons in strategic thinking and data preservation, emphasizing the importance of safeguarding data and effective collaboration. To address this issue, a standardized placeholder was created to depict missing units instead of discarding the data. This approach allows for easy querying and updating within the openEHR repository.

Addressing inconsistent decimal separators highlights the crucial balance between technical expertise and adaptability, emphasizing the need for cultural sensitivity in data representation. This experience underscores the importance of considering regional nuances and adapting technical solutions accordingly. These insights provide valuable guidance for maintaining data integrity across diverse contexts. Consistency in numeric value formatting, including decimal separators, is vital for ensuring data correctness and facilitating accurate interpretation for clinical decision-making and research.

Encountering text-based values instead of numeric measurements in laboratory observations underscores the necessity for a comprehensive review of solution specifications tailored to the context. This highlights the critical importance of careful data specification reviews to ensure precise alignment with intended data formats and prevent deviations. Additionally, it sheds light on organizational deficiencies regarding the misuse of standards. For instance misusing archetypes and templates, especially in openEHR, can lead to discrepancies, as specific observations require unique templates designed specifically for those observations. e.g virology observations.

Similar to missing observation timestamps, duplicates in legacy data also stemed from the source systems. Thus, internal processes necessitate proactive solutions within the mapping process and advocacy for improved data practices at the source.

Collaboration emerged as a crucial aspect of the process, as it involved engaging domain experts to bridge technical and clinical perspectives. This collaboration was essential for formulating effective mapping solutions and preserving data authenticity. The iterative methodology allowed for continuous improvement of the ETL jobs, addressing discrepancies in legacy HL7 v2 data and enhancing mapping accuracy. Challenges exceeding the scope of our ETL jobs were reported back to the staff in charge of the source systems for resolution. This methodology enabled thorough exploration of data integration challenges and informs standards adherence, facilitating the creation of more precise clinical data repositories which are the basis of data-based research like the UKSH MeDIC.

5. Conclusions

To conclude, implementing robust data governance measures and standardizing data formats are essential for enhancing clinical data quality. These practices ensure consistency, completeness, and reliability, which are crucial for accurate clinical decision-making, improved patient care, and meaningful medical research. Collaboration between domain experts and technical professionals further strengthens data integration, leading to precise clinical data repositories. Iterative improvement of ETL jobs and proactive reporting of challenges to source system staff contribute to long-term data quality enhancement by addressing issues like missing timestamps, measurement units, and inconsistent decimal separators.

The success of initiatives like the UKSH MeDIC highlights the benefits of vigilant data management and collaboration. Ensuring data quality and interoperability supports clinical trials, epidemiological studies, and healthcare analytics, driving advancements in digital health. Comprehensive data governance measures empower healthcare professionals with reliable information, improving patient outcomes, advancing medical research, and optimizing healthcare operations. These efforts lay a strong foundation for future advancements in healthcare delivery and research, demonstrating the transformative potential of integrated clinical data.

Acknowledgements

We thank our Refinement team whose work has been exceptional in the mapping process of HL7 v2 data to openEHR. This work was funded by the Federal Ministry of Education and Research (grant number 01KX2121).

References

[1] Kraus S, Schiavone F, Pluzhnikova A, and Invernizzi AC, Digital transformation in healthcare: Analyzing the current state-of-research. Journal of Business Research. 2021;2023:557–567. doi:10.1016/j.jbusres.2020.10.030.

[2] Data Standardization – OHDSI, (n.d.). https://www.ohdsi.org/data-standardization/ (accessed March 6, 2024).

[3] Min L, Liu J, Lu X, Duan H, and Qiao Q, An Implementation of Clinical Data Repository with openEHR Approach, Studies in Health Technology and Informatics. 2016;227:100–105.

[4] Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, Knaup-Gregori P, Bavendiek U, Dieterich C, Brors B, Kraus I, Thoms CM, Jager D, Ellenrieder V, Bergh B, Yahyapour R, Eils R, Consortium H, and Marschollek M, HiGHmed - An Open Platform Approach to Enhance Care and Research across Institutional Boundaries. Methods Inf Med. 2018;57:e66–e81. doi:10.3414/ME18-02-0002.

[5] Cheng KY, Pazmino S, and Schreiweis B, ETL Processes for Integrating Healthcare Data - Tools and Architecture Patterns, Stud Health Technol Inform. 2022;299: 151–156. doi:10.3233/SHTI220974.

[6] Syed R, Eden R, Makasi T, Chukwudi I, Mamudu A, Kamalpour M, Kapugama Geeganage D, Sadeghianasl S, Leemans SJJ, Goel K, Andrews R, Wynn MT, ter Hofstede A, and Myers T, Digital Health Data Quality Issues: Systematic Review, J Med Internet Res. 2023;25:e42615. doi:10.2196/42615.

[7] Reimer AP, and Madigan EA, Veracity in big data: How good is good enough. Health Informatics J. 2019;25:1290–1298. doi:10.1177/1460458217744369.

[8] Poppe E, Pika A, Wynn MT, Eden R, Andrews R, and ter Hofstede AHM, Extracting Best-Practice Using Mixed-Methods, Bus Inf Syst Eng. 2021;63:637–651. doi:10.1007/s12599-021-00698-9.