Healthcare Transformation with Informatics and Artificial Intelligence
J. Mantas et al. (Eds.)

doi:10.3233/SHTI230496

Achieving Interoperable Datasets in Pediatrics: A Data Integration Approach

Louisa BODE[a,+,1] Marcel MAST[a,+], Henning RATHERT[b], ELISE STUDY GROUP[*], Thomas JACK[b] and Antje WULFF[a,c]
+ contributed equally
**ELISE STUDY GROUP members will be communicated separately*
[a] *Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany*
[b] *Department of Pediatric Cardiology and Intensive Care Medicine, Hannover Medical School, Hannover, Germany*
[c] *Big Data in Medicine, Carl von Ossietzky University Oldenburg, Oldenburg, Germany*

Abstract. Despite their increased secondary value for developing applications and knowledge gain, routine, harmonized and standardized datasets are often not available in Pediatrics. We propose a data integration pipeline towards an interoperable routine dataset in pediatric intensive care medicine. Our three-level approach involves identifying relevant data from primary source systems, developing local data integration processes, and converting data into a standardized, interoperable format using openEHR. We modeled 15 openEHR templates and established 31 interoperable ETL processes, resulting in anonymized, standardized data of about 4,200 pediatric patients that were loaded into a harmonized database. Based on our pipeline and templates, we successfully integrated the first part of this data in our openEHR data repository. We seek to inspire other pediatric intensive care units to adopt similar approaches, with the aim of breaking down heterogenous data silos and promoting secondary use of routine data.

Keywords. Pediatric Intensive Care Units, Dataset, Interoperability

1. Introduction

Comprehensive, longitudinal medical routine datasets are of great value beyond their primary purpose of documentation [1]. However, obtaining Electronic Health Record (EHR) data for secondary uses is difficult. Often, vendor-dependent, proprietary and heterogeneous data sources are installed and inherently attended by a lack of standardization. For valuable reuse, additional effort is required to harmonize and transform data into interoperable formats [1,2]. To ensure interoperability of medical data, standards as HL7 CDA/CCR, HL7 FHIR or openEHR are used along with ontologies or terminologies [3]. OpenEHR emerged to be a reliable, vendor-neutral data infrastructure for the storage, retrieval, and exchange of EHRs, using a two-level

[1] Corresponding Author: Louisa Bode, Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Karl-Wiechert-Allee 3, 30625 Hannover, Germany. Email: louisa.bode@plri.de.

methodology [4]. The reference model (first level) defines data structures and types, while archetypes and templates (second level) hold the actual health data [4]. Archetypes represent a single clinical concept, while templates group multiple archetypes into one for specific clinical scenarios. Both archetypes and templates are co-developed with domain experts and can be shared, reused, and customized, promoting interoperability and data consistency [3]. They also allow for standardized terminology bindings, reducing linguistic ambiguities.

In our previous work, we demonstrated that it is possible to build interoperable applications upon openEHR repositories, specifically Clinical Decision Support Systems for the detection of pediatric Systemic Inflammatory Response Syndrome (SIRS), sepsis, and associated organ dysfunctions [2,5]. Now, we strive for feeding the applications with larger datasets to optimize the underlying models and to reach an evolutionary open pediatric intensive care dataset (see [6]). To integrate enhanced data, we require an extensive data integration process. With this article, we present our data integration pipeline that paves the way to a standardized, longitudinal pediatric intensive care dataset.

2. Methods

The *first step* of our three-level approach (see Figure 1) for a data integration pipeline comprises identification of data assets highly relevant for pediatric intensive care. Both the primary source systems and the individual clinical parameters needed to be mined. To deal with the highly complex and variable nature of pediatric patient's underlying diseases and treatment pathways, we closely collaborated with clinical experts. Once relevant data was identified, we established data access by considering each system's specifics, as they vary in underlying database structures. The *second step* involves the development of local Extract, Transform, and Load (ETL) processes, harmonizing data in an interim database. Both clinical and technical members monitored the ETL process regularly to ensure its quality. Given the high variety in data formats and sources, harmonization was crucial to ensure consistency and standardization across different sources. Thus, in a *third step*, we cooperatively modeled openEHR templates and mapped the original database columns to the corresponding archetype paths (by using an open-source tool called HaMSTER [7]) to finally load data in a standardized format into an openEHR-based clinical data repository (EHRbase, available at www.github.com/ehrbase). If possible, we reused existing, internationally agreed-upon archetypes from the Clinical Knowledge Manager (CKM).

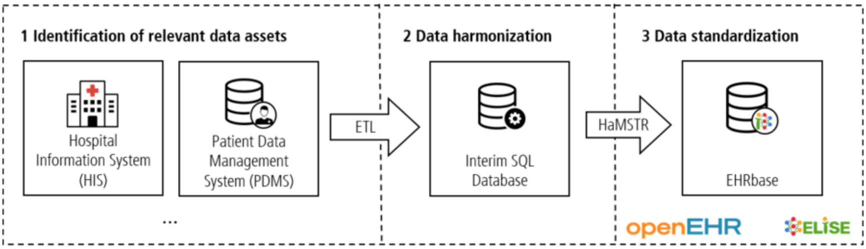
1 Identification of relevant data assets
Hospital Information System (HIS)
Patient Data Management System (PDMS)
...
ETL
2 Data harmonization
Interim SQL Database
HaMSTR
3 Data standardization
EHRbase
openEHR
ELISE

Figure 1. Three-level approach of data integration pipeline.

3. Results

First step. To ensure effective integration of a patient's EHR data, knowledge about the hospital's IT infrastructure and existing communication paths between systems was essential. Overall, more than 3000 data items (see Table 1) have been identified.

Second step. An interim SQL database structure and 31 subsequent ETL processes were created. In our final SQL database, we integrated an anonymized raw dataset of over 4,200 pediatric patients from the Pediatric Cardiology and Intensive Care Unit (PICU) of the Hannover Medical School (MHH). The dataset was further expanded by integrating data of these patients from twelve other pediatric MHH wards, resulting in a dataset of 250 million rows. Due to German data protection laws and our ethical responsibility to safeguard the privacy of patients featured in this dataset, data was anonymized by removing or replacing sensitive information while ensuring the data's accuracy and temporal sequencing are maintained.

Third step. We separated the tasks into a) modeling clinical concepts for the storage of patient data related to sepsis and SIRS and b) loading the data into EHRbase. For a), modeling activities have been guided by prior work, such as the FHIR representation of the German Medical Informatics Initiative's 'Intensive Care' core data set extension module or the existing templates in the context of intensive care hospitalizations in COVID-19 obtainable from the CKM. We reused approximately 45 existing archetypes and created 15 templates. The final templates were made publicly available in a CKM instance (https://ckm.highmed.org/ckm/projects/1246.152.38) and exported as Operational Template Format (OPT) files. For b), after installation of a local EHRbase repository, the exported OPT files were uploaded to EHRbase using the openEHR REST API specifications. Ultimately, the most intricate step was to map each data item in the database to the corresponding one in the template by utilizing the HaMSTR mapping tool. Once the mapping was finalized, openEHR compositions were sent to EHRbase for storage. From there, data can be accessed by the REST API.

Table 1. Overview of identified data items by category

Category	Approx. #Parameter	Category	Approx. #Parameter	Category	Approx. #Parameter
Patient-related	18	Fluid Balance	67	Therapy	15
Vital signs	13	Laboratory	2.911	Medication	6
Physical Status	6	Ventilation	136	External Material	37

4. Discussion

Interoperability standards such as openEHR have gained significant attention in recent years, resulting in increased research efforts and growth of documentation on standard-based data modeling and the use of interoperable data repositories such as EHRbase. By relying exclusively on open-source software (i.e., EHRbase, HaMSTR, Archetype Designer, and CKM) and making our data models openly available, we have created a data integration pipeline that can be reproduced by other researchers. However, extraction of data items according to their respective system's landscape remains a crucial, resource-intense step. Depending on the data available and their formats, templates may need to be adapted to meet local requirements. Moreover, terminology bindings were not yet considered in our models but will be incorporated in future revisions of the templates. Additionally, not all data has been transformed into the openEHR format to date, requiring further efforts. In the future, we strive to demonstrate the versatility and scalability of our data integration pipeline by also accommodating new data types and sources such as raw data signals like Electrocardiograms (ECGs), e.g., by using the ECG result archetype. This will be the next step towards our evolutionary, ever-growing dataset of critically ill children [6]. We are aware that anonymization datasets can still leak sensitive information, however, meticulous care was taken to ensure that any kind of sensitive information was removed, which e.g., also necessitated manual reviews of free texts. The details of our anonymization strategy and the final full data set itself are beyond the scope of this work and will be published separately.

We believe that our data integration pipeline will inspire other pediatric intensive care units to follow similar approaches, with the aim of breaking down data silos, overcoming interoperability issues and facilitating enhanced data-driven research in pediatric intensive care.

5. Conclusions

In this work, we proposed a three-level data integration pipeline that paves the way to a standardized, longitudinal and anonymized dataset in an interoperable format in pediatric intensive care.

Acknowledgments

This work is funded by the Federal Ministry of Health, Grant No. 2520DAT66A. Ethics approval for use of routine data was given by the Ethics Committee of Hannover Medical School (approval number 9819_BO_S_2021).

References

[1] Prince K, Jones M, Blackwell A, Simpson A, Meakins S, Vuylsteke A. Barriers to the secondary use of data in critical care. J Intensive Care Soc. 2018 May;19(2):127-31.

[2] Wulff A, Haarbrandt B, Tute E, Marschollek M, Beerbaum P, Jack T. An interoperable clinical decision-support system for early detection of SIRS in pediatric intensive care using openEHR. Artif Intell Med. 2018 Jul;89:10-23.

[3] Garde S, Knaup P, Hovenga E, Heard S. Towards semantic interoperability for electronic health records. Methods Inf Med. 2007;46(3):332-43.

[4] Beale T. Archetypes: Constraint-based Domain Models for Future-proof Information Systems. Paper presented at: Eleventh OOPSLA workshop on behavioral semantics; 2002:16–32.

[5] Bode L, Schamer S, Böhnke J, Bott OJ, Marschollek M, Jack T, Wulff A, ELISE Study Group. Tracing the Progression of Sepsis in Critically Ill Children: Clinical Decision Support for Detection of Hematologic Dysfunction. Appl Clin Inform. 2022 Oct;13(5):1002-1014.

[6] Wulff A, Mast M, Bode L, Rathert H, Jack T, Elise Study Group. Towards an Evolutionary Open Pediatric Intensive Care Dataset in the ELISE Project. Stud Health Technol Inform. 2022 Jun 29;295:100-103.

[7] Tute E. HAMSTRETLBuilder. Available from: https://gitlab.plri.de/tute/HAMSTRETLBuilder. Last access: March 20th 2023.