Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220670

Towards an Evolutionary Open Pediatric Intensive Care Dataset in the ELISE Project

Antje WULFF[a,b,1], Marcel MAST[a], Louisa BODE[a], Henning RATHERT[c], ELISE STUDY GROUP[*] and Thomas JACK[c]

**ELISE STUDY GROUP: Louisa Bode[a]; Marcel Mast[a]; Antje Wulff[a,b]; Michael Marschollek[a]; Sven Schamer[c]; Henning Rathert[c]; Thomas Jack[c]; Philipp Beerbaum[c]; Nicole Rübsamen[d]; Julia Böhnke[d]; André Karch[d]; Pronaya Prosun Das[e]; Lena Wiese[e]; Christian Groszewski-Anders[f]; Andreas Haller[f]; Torsten Frank[f]*

[a] *Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany*
[b] *Big Data in Medicine, Department of Health Services Research, School of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg, Oldenburg, Germany*
[c] *Department of Pediatric Cardiology and Intensive Care Medicine, Hannover Medical School, Hannover, Germany*
[d] *Institute of Epidemiology and Social Medicine, University of Muenster, Muenster, Germany*
[e] *Research Group Bioinformatics, Fraunhofer Institute for Toxicology and Experimental Medicine, Hannover, Germany*
[f] *Medisite GmbH, Hannover, Germany*

Abstract *Background:* To embrace the need for freely accessible training data sets originating from the real world, in the ELISE project, we integrate source data from a pediatric intensive care unit and provide it to researchers. *Objective:* We present our vision, initial results and steps on a trail towards an evolutionary open pediatric intensive care data set. *Methods:* Our evolution plan for the data set comprises three steps. The final data set will include raw clinical data and labels on critical outcomes such as organ dysfunction and sepsis, generated automatically by computerized and well-evaluated methods. *Results:* First step resulted in an initial version data set available in a central repository. *Conclusions:* Our approach has great potential to provide a comprehensive open intensive care data set labeled for critical pediatric outcomes and, thus, contributing to overcome the current lack of real-world pediatric intensive care data usable for training data-driven algorithms.

Keywords. Intensive Care Units, Pediatric; Dataset; Data Science

[1] Corresponding Author, Antje Wulff, Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Karl-Wiechert-Allee 3, 30625 Hannover, Germany; E-mail: antje.wulff@plri.de

1. Introduction

In medical data science, there is a recent shift from knowledge-based approaches towards the application of data-driven methods [1]. Taking the enormous growth of data and the numerous potentials of data-driven learning into account, this progression is advantageous, especially if clinically relevant findings can be derived. Thus, such growing databases could represent an enormous leap in the development towards evidence-based medicine. However, it is not possible for every data scientist to connect to real-world medical centers to access data. Furthermore, '*multiple use of routine data'* in the aforementioned sense is not straightforward since it requires time-consuming assembling and harmonization of heterogeneous and multi-source data [2]. If data are available and integrated, it would be fair to make it accessible for other researchers. The MIMIC database is an outstanding example for such an endeavor as it contains de-identified health-related data associated with over forty thousand patients from a critical care unit in Israel [3]. Another related database is eICU [4], also containing data of adults or neonates, but no data on critically ill children. Thus, in 2020, Zeng et al. published PIC, a pediatric-specific intensive care database comprising children's data [5].

In our project ELISE (a learning, interoperable and smart expert system for pediatric intensive care), we explore options on computerized, data-driven detection of pediatric systemic inflammatory response syndrome (SIRS), sepsis and organ dysfunction [6]. We strive for setting up processes to continuously assemble and standardize routine data from a patient data management system of the Pediatric Cardiology and Intensive Care Unit of the Hannover Medical School. Afterwards, we develop a knowledge-based approach using international diagnostic criteria as a basis (see Wulff et al. [7, 8]), evaluate its diagnostic accuracy and use this approach as a labeling mechanism for assigning outcomes labels to all patients retrospectively. Based on this training data, data-driven approaches for an early detection of the aforementioned diseases will be developed and integrated into an open demonstrator of a clinical decision-support system. For this task, it is not possible to use the PIC database [5] since it only contains diagnosis codes without exact time mappings of those critical events. Our ELISE dataset contains this highly important time resolution, creating an optimal structure for training data for machine learning. Our vision is to complement available databases by providing a pediatric intensive care data set with outcome labels, generated by well-evaluated computerized models. The data set will develop in an evolutionary way twofold: (1) in terms of scope, since we plan to add new parameters over time, (2) in terms of size, since we set up automatic local processes to continuously add new patient data for already integrated parameters. With this article, we communicate our vision as well as first results and steps on our trail towards an open pediatric intensive care data set.

2. Methods

The evolution of the data set is planned in three major steps (see Figure 1). *First*, a basic data set containing general patient and intensive care data such as vital signs series, laboratory values, device-related data, diagnoses, and procedure codes will be released. *Second*, outcome labels for SIRS and sepsis as well as hepatic, hematologic, renal, cardiovascular, and respiratory organ dysfunctions will be added. *Third*, raw data from electrocardiograms and related devices will be included. Each major step will be developed evolutionary itself in minor steps, e.g., for the basic data set, we first include only data from one ward, prospectively enhancing with data from other wards. This can provide a more complete picture of the patient's entire hospital stay. We set up local extraction, load and transform processes (ETL) using Microsoft SQL Server Integration Services to gather data from the primary source systems. ETL processes can be carried out continuously to be able to add new patient data, and data is automatically de-identified. Transformed data are stored in a Microsoft SQL staging database and, later, by using the self-developed HaMSTR tool [9] in a standardized semantically enriched data platform using openEHR [10] or HL7 FHIR [11]. Add-ons will be delivered consecutively, comprising (a) data pre-processing packages, (b) labeling algorithms used, (c) standard-based representation of our data, (d) our open demonstrator app for data-driven detection of the above-mentioned conditions.

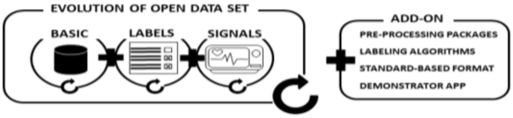

Figure 1. The evolution of the open pediatric intensive care data set and its add-ons.

3. Results

The basic dataset (see Table 1) with a data description is available in csv format in a central data repository, accessible upon request [12]. For the second major release, labels for hematologic and hepatic dysfunctions are already prepared and available with all other labels in the second version of the data set by the end of the year. To expand raw data towards the third release, we currently work on integrating data from the central monitoring system. The aim is to include vector-based data and curves, e.g., for arterial and central venous pressures and the plethysmographic function for pulse oximetry.

Table 1. Overview of the scope of the first basic dataset (seven years of MHH PICU data)

Parameter	Approx. Entries	Parameter	Approx. Entries
Patients	5.500	Medications	4.000.000
Vital signs	90.000.000	Procedures	400.000
Device-related data	20.000.000	Diagnoses	200.000
Laboratory parameter	8.000.000	Movement data	100.000

4. Discussion

We experienced the extraction of primary source data as a time-consuming task requiring enormous initial efforts. Due to heterogeneous data representations, interpretation of data required close collaboration with clinicians, manufacturers, and system administrators. Compared to related studies, such as the MIMIC [3] or PIC [5] studies, our current data set is limited in matters of size and variety. However, we are confident that our evolution plan allows delivering a similar comprehensive data set within the time span of ELISE. Additionally, we strongly concentrate on contributing outcome labels as discussed, which are not yet available in the above-mentioned related initiatives and databases. In contrast to the other studies, we strongly focus on international data standardization and classification (e.g. by using openEHR [10], FHIR [11], SNOMED [13], LOINC [14]) to exculpate data scientists from data preprocessing. This also will facilitate multi-center-approaches, thus, reaching a new level of quality in data science. In all matters, data security is a risk and needs to be considered by both data providers and data users. Our anonymization approach comprises de-identification. Currently, we work on gathering requirements on new data to be added in future major versions, e.g., from Anesthesia. With this publication, we communicate our approach in an early manner to be able to collect further requirements from interested readers, researches and data scientists.

5. Conclusions

By developing a comprehensive open intensive care data set, labeled for critical outcomes, our work contributes to overcome the current lack of available pediatric intensive care data reusable to train data-driven algorithms. Further input and requirements from the international community are much appreciated.

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] Chang AC. Intelligence-based medicine: Artificial intelligence and human cognition in clinical medicine and healthcare. AIMed Artificial intelligence in medicine. London, United Kingdom, San Diego, CA, United States: Elsevier Academic Press; 2020. 521 p.

[2] 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.

[3] Johnson A, Bulgarelli L, Pollard T, Horng S, Celi L A, Mark R. MIMIC-IV (version 1.0). PhysioNet. 2021. Available from: https://doi.org/10.13026/s6n6-xd98, Last access: 14 March 2022.

[4] Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018 Sep;5:180178.

[5] Zeng X, Yu G, Lu Y, Tan L, Wu X, Shi S, et al. PIC, a paediatric-specific intensive care database. Sci Data. 2020 Jan;7(1):14.

[6] Peter L. Reichertz Institute for Medical Informatics. PLRI | ELISE. Available from: https://plri.de/en/forschung/projekte/elise, Last access: 14 March 2022.

[7] Wulff A, et al. 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.

[8] Wulff A, Montag S, Rübsamen N, Dziuba F, Marschollek M, Beerbaum P, et al. Clinical evaluation of an interoperable clinical decision-support system for the detection of systemic inflammatory response syndrome in critically ill children. BMC Med Inform Decis Mak. 2021 Feb;21(1):62.

[9] Tute E. HAMSTRETLBuilder. Available from: https://gitlab.plri.de/tute/HAMSTRETLBuilder. Last access: 25 April 2022.

[10] T. Beale, in Eleventh OOPSLA workshop on behavioral semantics: serving the customer, Northeastern University, Seattle, Washington, Boston, 2002, 16–32.

[11] HL7. FHIR v4.0.1, 2019;10. Available from: http:// hl7.org/fhir, Last access: 14 March 2022.

[12] Wulff A, Mast M, Bode L, Rathert H, Jack T, Marschollek M, et al. ELISE - An open pediatric intensive care data set [Data set]. Available at: https://publikationsserver.tu-braunschweig.de/receive/dbbs_mods_00070468, Last access: 14 March 2022.

[13] Regenstrief Institute. LOINC. Available from: https://loinc.org/. Last access: 25 April 2022.

[14] SNOMED International. SNOMED. Available from: http://www.snomed.org/. Last access: 25 April 2022.