A Pipeline for the Usage of the Core Data Set of the Medical Informatics Initiative for Process Mining - A Technical Case Report

Hauke HEIDEMEYER[a,b,c], Leo AUHAGEN[a,b,c], Raphael W. MAJEED [a,d], Marco PEGORARO[c], Jonas BIENZEISLER[a], Viki PEEVA[c], Harry BEYEL[c], Rainer RÖHRIG[a], Wil M. P. VAN DER AALST[c], and Behrus PULADI[a,b,1]

[a] *Institute of Medical Informatics, University Hospital RWTH Aachen, Aachen, Germany*

[b] *Department of Oral and Maxillofacial Surgery, University Hospital RWTH Aachen, Aachen, Germany*

[c] *Chair of Process and Data Science, RWTH Aachen University, Aachen, Germany*

[d] *Universities of Giessen and Marburg Lung Center (UGMLC), Member of the German Center for Lung Research (DZL), Giessen, Germany*

ORCID: Hauke Heidemeyer https://orcid.org/0000-0002-7401-6860

Abstract. Introduction: Process Mining (PM) has emerged as a transformative tool in healthcare, facilitating the enhancement of process models and predicting potential anomalies. However, the widespread application of PM in healthcare is hindered by the lack of structured event logs and specific data privacy regulations. **Concept:** This paper introduces a pipeline that converts routine healthcare data into PM-compatible event logs, leveraging the newly available permissions under the Health Data Utilization Act to use healthcare data. **Implementation:** Our system exploits the Core Data Sets (CDS) provided by Data Integration Centers (DICs). It involves converting routine data into Fast Healthcare Interoperable Resources (FHIR), storing it locally, and subsequently transforming it into standardized PM event logs through FHIR queries applicable on any DIC. This facilitates the extraction of detailed, actionable insights across various healthcare settings without altering existing DIC infrastructures. **Lessons Learned:** Challenges encountered include handling the variability and quality of data, and overcoming network and computational constraints. Our pipeline demonstrates how PM can be applied even in complex systems like healthcare, by allowing for a standardized yet flexible analysis pipeline which is widely applicable.The successful application emphasize the critical role of tailored event log generation and data querying capabilities in enabling effective PM applications, thus enabling evidence-based improvements in healthcare processes.

Keywords. Process Mining, Health Information Systems, Data Integration, FHIR, Healthcare Quality Assurance, Event Log

1. Introduction

1.1. Background

Process Mining (PM) is the science of utilizing log data to: 1.) Discovery of existing processes; 2.) Conformance check if processes fits created models; 3.) Enhancement of existing process models. Using PM upcoming anomalies can be predicted and proactively influence parts of the problematic processes [1]. Thereby, PM could improve healthcare efficiency and patient outcomes through systematic process analysis in the future [2–4].

PM is widely used in manufacturing and service organizations to improve process efficiency, transparency, and flexibility [1]. In healthcare, it is not widely applied beyond single case studies within a research context [2]. This is partly because PM analyses necessitate a clean, structured event log, which is not immediately available in a hospital's information system. The challenges are the high variability and heterogeneity of behaviors in healthcare processes, data from varying abstraction levels, generally low data quality, and the need for multidisciplinary domain knowledge [2,3]. To make matters worse, the lack of established pipelines and the high complexity hinder the application of PM to healthcare with regard to more traditional field of applications [2]. In this regard, the German healthcare system faces significant challenges in adopting new technologies due to the lack of standardized routine data [5,6].

However, healthcare data processors can access a standardized Core Data Set (CDS) defined by the Medical Informatics Initiative (MII). This data is provided by respective Data Integration Centers (DICs) by all German university hospitals. The CDS data comprises seven basic and eleven extension modules, each tailored with Fast Healthcare Interoperable Resources (FHIR) specifications. Basic modules are universally applicable, like the patient module, while extension modules target specific disciplines such as oncology. This ensures standardized and harmonized data across healthcare settings [7]. Unfortunately, this data is not in an event log format, which is necessary for PM [8].

Furthermore, until now only anonymized data or data based on broad consent from DICs could be shared among researchers due to data privacy requirements. These are all challenges for the application of PM in health care. Nevertheless, with the recent Health Data Utilization Act (HDUA) a consent-free sharing of healthcare routine data for quality assurance, patient safety, and medical research within a publicly funded consortium of healthcare data processors is possible [9,10]. Since HDUA is no longer limited to patient consent, a large data set would now be available for PM from DICs or non-university hospitals. For PM to be widely applicable across hospitals, it must be possible to convert routine data into event logs, regardless of their source. This should take into account existing standards such as the CDS whose data is commonly stored in the FHIR format. But also, the use of non-FHIR data should be allowed like from non-university hospitals.

In this regard, we present a pipeline that supports the widespread use of PM in respect of the CDS of the MII. The pipeline converts routine data into Fast Healthcare Interoperable Resource (FHIR) records, stores them in a local FHIR server, and then converts them into a PM event log standard via FHIR queries.

1.2. Objective and Requirements

Our primary objective is to enable the wide application of PM based on the CDS available of the DICs. The pipeline to this objective should fulfil the following listed requirements:

R1) The pipeline must utilize the CDS as basis for event log creation.
R2) The pipeline must be compatible with current MII standards.
R3) The abstraction level of the event log must be adjustable.
R4) The pipeline should be extensible to non-university data.

Limiting the system to the CDS (R1) restricts the range of extractable events from the data but allows for predefined event extraction methods. The system must also integrate seamlessly with any DIC (R2), ensuring broad usability. R3 ensures the event log supports various research needs at different abstraction levels. Finally, R4 enables non-university hospitals without a DIC to integrate their data into our pipeline.

2. State of the art

2.1. Related Work

In PM, data standards for event logs like the IEEE-specified eXtensible Event Stream (XES) [11] have been recently augmented by the Object Centric Event Log (OCEL) [12] format. OCEL enables PM to analyze relationships between different objects within processes which offers a more comprehensive view of processes [13,14].

Previous work mainly focused on the utilization and creation of audit events for event extraction from routine data [4,14–18] and subsequent process improvement based on the valuable insights healthcare routine data can provide [19,20].

Cruz-Correia et al. established the connection between automatically generated audit events by the HIS and their application for PM [15]. Subsequently, Helm et al. expanded this approach to generate an standardized event log based on events created [17]. González López de Murillas et al. focused on creating an event log that contains as much information as possible about different objects, such as patients and physicians, within the event log [16]. Gatta et al. developed a methodical approach to create an event log using the HL7 FHIR Audit Event resource includes relevant operational, security, and privacy events [18]. Helm et al. advanced the approach by modifying a FHIR server to intercept changes to resources, generating audit event resources that are then converted into event logs in the XES format [4]. Finally, Cruz-Correia et al. utilized the OCEL event log for their audit event log extraction [15].

The development of a structured event log extraction framework, as demonstrated for the MIMIC data set, has already shown significant utility in the healthcare community for broader applications that require systematic and standardized data analysis [21].

2.2. Shortcomings

While there exist approaches to provide frameworks or methodologies for the event extraction based on standardized data sources, to the best of our knowledge there exists no system based on the CDS provided by the DICs. Existing generalizable system utilize the audit event resource, not included in the MII CDSs' specification and requiring existing changes to the FHIR servers' implementation, violating R1-R2. Additionally, the transformation of Helm et al. to XES includes during the implementation of the audit resource generation a specific event abstraction level, with different level requiring a change to the existing implementation. Consequently, all existing methodologies are partly in conflict with R1-R4, requiring a custom system explained in the following.

3. Concept

In order to simulate the complete data flow from the routine data to the event log we used a raw data set from the Oral Maxillofacial Surgical (OMFS) department at University Hospital RWTH Aachen (UHRA) [22].

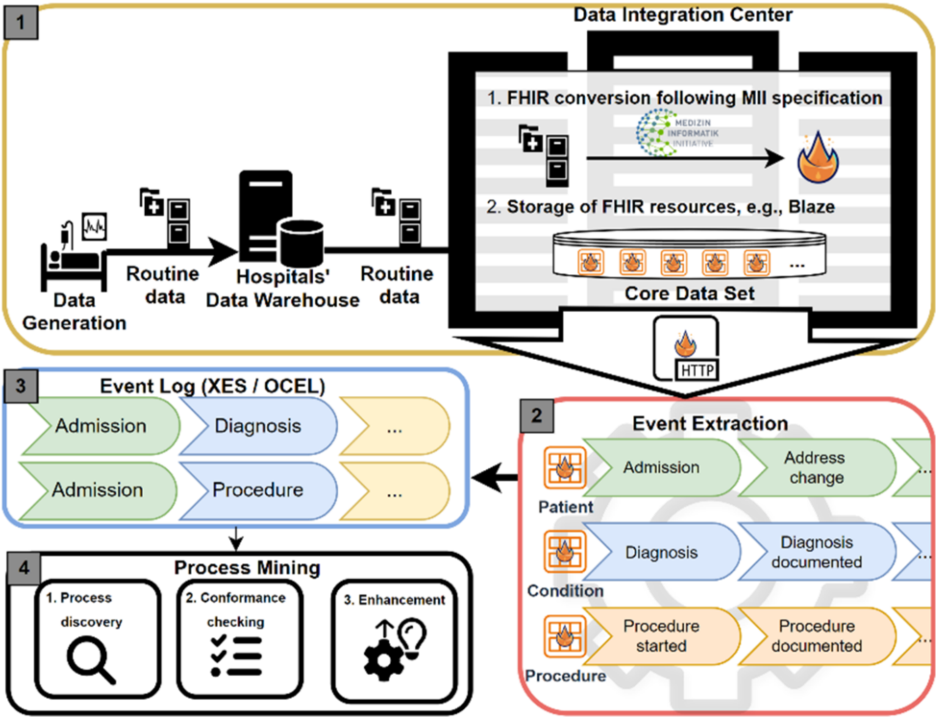

Figure 1. The conceptual architecture consists out of three main parts. The first part, highlighted in yellow, details the flow of routine data from its electronic transmission to the hospital data warehouse, accessible by various hospital information systems, and its subsequent transfer to the Data Integration Center (DIC). Here, relevant information is extracted from the Core Data Set (CDS) as specified by the Medical Informatics Initiative and stored on a Fast Healthcare Interoperable Resources (FHIR) data store, where it is represented by FHIR resources. We simulated a DIC using routine data from the hospital's data warehouse, which was then extracted and transformed into the FHIR CDS format, and stored in Blaze, a FHIR data store. This FHIR data store allows resource access via HTTP due to the FHIR specification. The second part, indicated by a red box (step 2), involves creating a system that extracts events from the FHIR resources provided by *Blaze*, including patient, condition, procedure, and case resources. The third part (blue box, step 3) transforms these events into an event log, which facilitates the use of Process Mining to identify, validate, and enhance process models.

In future, hospitals without DICs could implement our method by converting data from their data warehouses into CDS-compliant data sets in FHIR format or by using existing FHIR representations to meet these specifications. To ensure a consistent and validated representation of UHRA's DIC in line with the FHIR specification and functional requirements, *Blaze*, a Docker-based FHIR store, has been deployed. Widely used among DICs for its CDS compliance, *Blaze* was preferred over other FHIR stores. Utilizing the FHIR API, events are extracted, an event log is created, and PM is performed.

In terms of methodology of event extraction, we adhere closely to the approach proposed by Helm et al., although we do not rely on a modified FHIR data store. Instead, we make use of the immutable nature of FHIR resources, which results in the creation of a new instance of the resource with the updated values and the version identifier increased by one. As *Blaze* adheres to the FHIR specification, it is possible to query all resources as well as their different versions via API. The creation and update of events can be performed for all resources and their corresponding fields without the necessity to modify the DICs' FHIR store in accordance with the requirements R1-R2. The generated events are stored in a standardized event log. This event log can then be used with tools such as *PM4Py* and ProM[2] for PM.

4. Implementation

4.1. Architecture

The programming language Python is used for the implementation as it is a popular choice for data science as well as process mining, with popular libraries for PM such as *PM4Py*[3] already available [23]. For the transfer of patients' data to the locally deployed *Blaze* instance, the package *fhir.resources*[4] has been extended to adhere to the MII CDSs' specification, with the advantage of the packages' implementation to verify the FHIR generated resources adhering to the CDSs' specification.

For communication with the FHIR data store, in our case Blaze, we use the standard library *requests*. We also used the *blazectl*[5] command line utility for FHIR resource extraction out of Blaze in the *ndjson* format for performance reasons. Additionally, we leverage the hierarchical structure of ICD-10 and OPS codes to achieve different event granularity.

Each modification of a resource generates a distinct event for precise logging, covering updates like patient addresses and symptom onset with varying detail, including city changes. References to related resources are maintained in the event log, enabling comprehensive process analysis across various entities, including procedures, diagnoses, and patients.

4.2. Solution and Results

We conducted a case study with all patients on the OMFS from 2011 to 2022, including 137,555 patients. We used the admission of patients, their diagnosis events and

[2] https://github.com/promworkbench

[3] https://github.com/pm4py

[4] https://github.com/glichtner/fhir.resources

[5] https://github.com/samply/blazectl

procedures as events on the highest abstraction level for simplicity in the visualizations. From 137,555 traces we discovered 4,273 variants including 719,025 activities. Most traces include just the patient registrations with no subsequent associated event, possibly suggesting an inclusion of outpatients of the OMFS without reporting diagnosis and procedures in ICD10 and OPS format and thus not captured by the extracted events (cf. figure 2).

Despite the Petri net from the inductive miner achieving an average fitness of about 0.73 (see figure 3), it offers limited process insights due to the highly parallel nature of the activities. A dotted chart in *ProM* showed no significant impact of the COVID-19 pandemic on the procedures. Only the procedure medication administration appears more often; however, due to the small number of events (23 starting on February 2019 until 2022, 41 in the entire time range) this observation might not be significant.

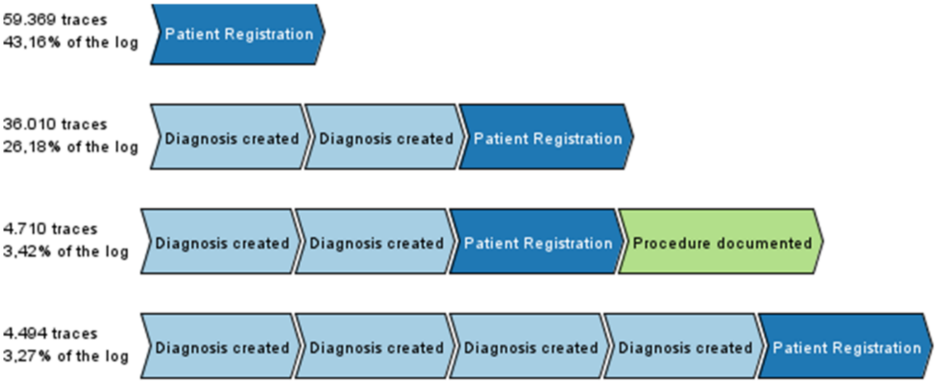

Figure 2. The figure visualizes common trace variants using diagnosis events at the highest abstraction level and procedures by the first letter of the OPS code created in *ProM*. Notably, many traces only contain patient registrations without subsequent events. This suggests an inclusion of outpatients of the OMFS without recorded diagnosis and procedures in ICD10 and OPS format and thus not captured by the extracted events. Approximately 3% of traces show two diagnoses before registration, and another 3% depict a sequence where a diagnosis is followed by patient registration and one operation procedure.

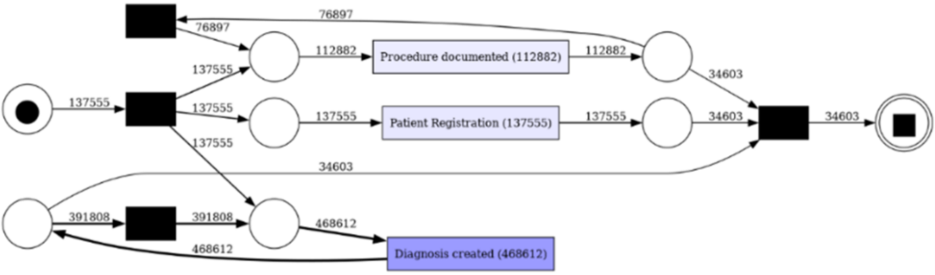

Figure 3. Petri net created using the Inductive Miner in *PM4Py* with a noise threshold of 30%. It is apparent that the activities are highly parallelized. The conformonce check of the log on the Petri net using *PM4Py* resulted in an average fitness of ca. 0.73 (0 worst fitness, 1 best).

For a more fine-grained analysis, we considered all patient hospital visits in the year 2020 (during the COVID-19 pandemic) and compared them to those in 2018 (before the pandemic started). From the OPS codes, we considered the first letter as representing the general procedure category. For diagnoses, we analyzed the first three letters of their corresponding ICD-10-GM codes in addition to the official registration of the patient case as a separate event. We filtered the 249 traces for the year 2020 by selecting all visits that included at least one diagnosis of Neoplasms (ICD-10-GM codes C00-D48). We then selected the top 100 variants and created a performance Directly Follows Graph (DFG), removing all edges with an absolute frequency of less than 10 to remove noise. Subsequently, all remaining activities not reachable from the start or end were removed to reduce the DFG (see figure 4). The same analysis was performed for the 225 traces from the year 2015 (see figure 5). It is evident that the process is different between the years. For example, the time of patient case registration to the diagnose C04 is substantially longer in the year 2020 than in the year 2018.

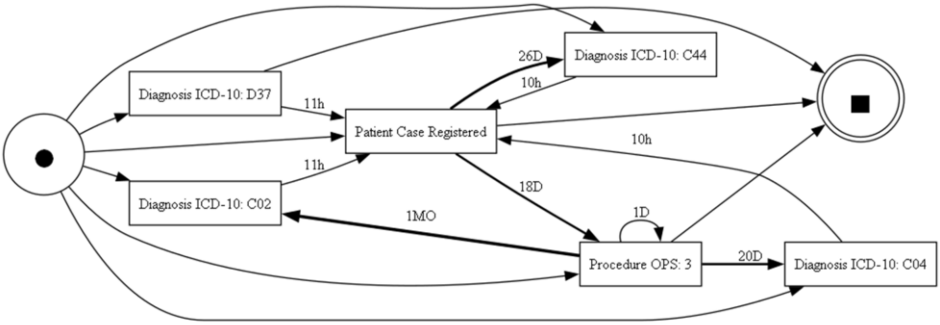

Figure 4. The reduced performance Directly Follows Graph for the filtered traces of the year 2020 annotated with the mean time between different events. The alignment based replay fitness of the DFG has an average fitness of ca. 0.98 (0 worst fitness, 1 best).

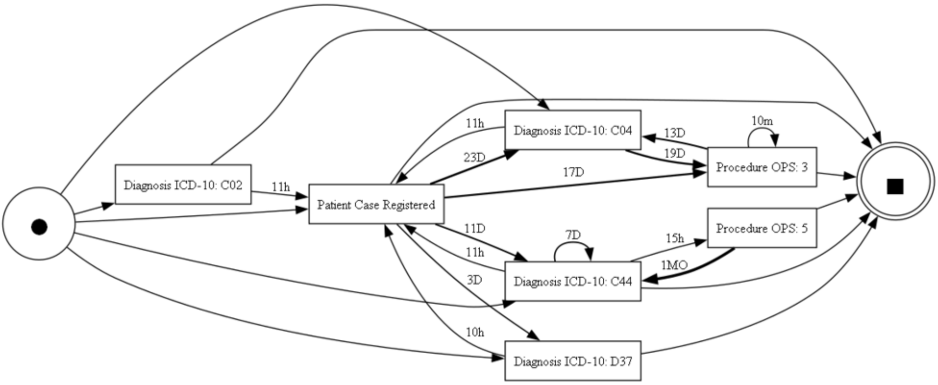

Figure 5. The reduced performance Directly Follows Graph for the filtered traces of the year 2018 annotated with the mean time between different events. The alignment based replay fitness of the DFG has an average fitness of ca. 0.98 (0 worst fitness, 1 best).

Even though DFGs allow for a simple process representation, the identification of bottlenecks and process inefficiencies is not possible with such a simple representation. For more meaningful insights, more advanced process mining algorithms in conjunction with extensive data cleaning are required, which is beyond the scope of this paper.

5. Lessons learned

Our methodology enables PM on routine data from university hospitals in Germany or hospitals' DICs following MIIs' specification of the CDS without altering existing DICs. While the FHIR API facilitates data querying, it poses computational and network resource challenges, albeit with easier integration. Although FHIR offers several advantages, its usage lacks performance guarantees compared to traditional methods like SQL, particularly evident in data transfer bottlenecks via HTTP. A bulk query with FHIR resulting in data in the *ndjson* format is favorable but is depended on the support of the FHIR store as it is an optional feature as well as FHIR transaction bundles. In the case of support for such features, the performance would benefit greatly by no longer requiring single HTTP requests.

Restricting event extraction based on data from the CDS ensures guaranteed event creation but represents only a subset of routine data, limiting fine-grained process analyses. Using the audit event resource in FHIR could effectively represent events within its framework but may conflict with resources *like ProcedureRequest* and *ServiceRequest*, risking ambiguous and inconsistent data representations.

Given the focus on CDS for wide applicability, custom solutions may be needed for hospital-specific analyses focusing on processes not represented by the CDS. Another challenge is the extraction of relevant process information from generated events due to inconsistencies in routine data documentation and the general low quality of routine data in healthcare. This hampers the insights when employing PM for inter-hospital analysis. Routine data also includes artifacts arising from Extraction, Transfer, Loading (ETL) processes and billing systems, which can further obscure real processes. This necessitates extensive effort in identifying subprocesses and filtering events whose application has been shown in recent research.

Despite these challenges, it is still possible to extract useful process models from routine data. In contrast to the simplified process model presented in this paper, a more fine-grained analysis of subprocesses is possible and is planned. The extraction of useful process models is, however, challenging and requires robust methodologies to handle data inconsistencies and variability which is planned in the future. With modern techniques and tools, even the highly complex and low-quality data found in healthcare settings can yield valuable insights. Consequently, this pipeline is only one important part of the many required for the successful application of PM on routine data.

6. Conclusion

- Our pipeline enables PM on routine data from university hospitals in Germany or hospitals' DICs following MIIs' CDS specification, without altering existing DICs.
- The FHIR API facilitates data querying, but poses computational and network resource challenges, lacking the performance guarantees of traditional methods like SQL.
- While restricting event extraction to CDS data ensures event creation, it limits the scope of process analyses, necessitating custom solutions for comprehensive insights.

- Routine data quality and artifacts from ETL processes and billing systems obscure real processes, requiring extensive effort in identifying and filtering relevant subprocesses.

Declarations

Ethical vote: Independend Ethical Committee of the medical faculty of the RWTH Aachen University, chair: Prof. Dr. Hausmann, vote-no: EK 23-086, 06.04.2023.

Funding: RM received funding from Federal Ministry of Education and Research (BMBF) for the development of the German Health Research Data Portal (FDPG-PLUS, 01ZZ2309E). Behrus Puladi was supported by the Clinician Scientist Program of the Faculty of Medicine RWTH Aachen University.

Conflict of Interest: The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: WvdA has an affiliation with the commercial software vendor Celonis SE. However, this affiliation did not influence this work in any way or affect objectivity. The other authors declare no conflict of interest according to ICMJE recommendations.

Contributions of the authors: HH, BP, RM, RR and WvdA conception of the work. HH implementation. BP and RM code review. HH writing the manuscript. BP, HH, LA, RM, MP, JB, VP, HB, RR and WvdA critical review, analysis and interpetation. All authors approved the manuscript in the submitted version and take responsibility for the scientific integrity of the work. This work is a part of the Master Thesis of Hauke Heidemeyer.

Acknowledgement: We thank Mr. Albert Esser for the data extraction.

REFERENCES

[1] W.M.P. van der Aalst, Process Mining: Data Science in Action, Springer Berlin Heidelberg, Berlin, Heidelberg, 2016. doi:10.1007/978-3-662-49851-4.

[2] J. Munoz-Gama, N. Martin, C. Fernandez-Llatas et al., Process mining for healthcare: Characteristics and challenges. J Biomed Inform 127 (2022), 103994. doi:10.1016/j.jbi.2022.103994.

[3] R.S. Mans, W.M.P. van der Aalst and R.J.B. Vanwersch, Process Mining in Healthcare, Springer International Publishing, Cham, 2015. doi:10.1007/978-3-319-16071-9.

[4] E. Helm, O. Krauss, A. Lin et al., Process Mining on FHIR - An Open Standards-Based Process Analytics Approach for Healthcare. In Process Mining Workshops, S. Leemans and H. Leopold, eds. Springer International Publishing, Cham, 2021, pp. 343–355. doi:10.1007/978-3-030-72693-5_26.

[5] G. Gopal, C. Suter-Crazzolara, L. Toldo et al., Digital transformation in healthcare - architectures of present and future information technologies. Clinical Chemistry and Laboratory Medicine (CCLM) 57 (2019), 328–335. doi:10.1515/cclm-2018-0658.

[6] J. Hauswaldt, V. Kempter, W. Himmel et al., Hindernisse bei der sekundären Nutzung hausärztlicher Routinedaten. Gesundheitswesen 80 (2018), 987–993. doi:10.1055/a-0668-5817.

[7] D. Ammon, M. Kurscheidt, K. Buckow et al., Arbeitsgruppe Interoperabilität: Kerndatensatz und Informationssysteme für Integration und Austausch von Daten in der Medizininformatik-Initiative. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 67 (2024), 656–667. doi:10.1007/s00103-024-03888-4.

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

[9] Bundesgesetzblatt Teil I - Gesetz zur verbesserten Nutzung von Gesundheitsdaten: GDNG. In Bundesgesetzblatt, 20/04/2024.

[10] J. Schmitt, T. Bierbaum, M. Geraedts et al., Das Gesundheitsdatennutzungsgesetz – Potenzial für eine bessere Forschung und Gesundheitsversorgung. Gesundheitswesen 85 (2023), 215–222. doi:10.1055/a-2050-0429.

[11] IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams, IEEE, Piscataway, NJ, USA. doi:10.1109/IEEESTD.2016.7740858.

[12] A.F. Ghahfarokhi, G. Park, A. Berti et al., OCEL: A Standard for Object-Centric Event Logs. In New Trends in Database and Information Systems Springer International Publishing, Cham, 2021, pp. 169–175. doi:10.1007/978-3-030-85082-1_16.

[13] W.M.P. van der Aalst, Object-Centric Process Mining: Dealing with Divergence and Convergence in Event Data. In Software Engineering and Formal Methods, P. C. Ölveczky and G. Salaün, eds. Springer International Publishing, Cham, 2019, pp. 3–25. doi:10.1007/978-3-030-30446-1_1.

[14] A. Pointner, O. Krauss, A. Erhard et al., Multi-Perspective Process Mining Interfaces for HL7 AuditEvent Repositories: XES and OCEL. Stud Health Technol Inform 301 (2023), 168–173. doi:10.3233/SHTI230034.

[15] R. Cruz-Correia, I. Boldt, L. Lapão et al., Analysis of the quality of hospital information systems Audit Trails. BMC Med Inform Decis Mak 13 (2013), 84. doi:10.1186/1472-6947-13-84.

[16] E. González López de Murillas, E. Helm, H.A. Reijers et al., Audit Trails in OpenSLEX: Paving the Road for Process Mining in Healthcare. In Information Technology in Bio- and Medical Informatics, M. Bursa, A. Holzinger, and M. E. Renda, eds. Springer International Publishing, Cham, 2017, pp. 82–91. doi:10.1007/978-3-319-64265-9_7.

[17] E. Helm and F. Paster, First Steps Towards Process Mining in Distributed Health Information Systems. International Journal of Electronics and Telecommunications 61 (2015), 137–142. doi:10.1515/eletel-2015-0017.

[18] R. Gatta, M. Vallati, J. Lenkowicz et al., A Framework for Event Log Generation and Knowledge Representation for Process Mining in Healthcare. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)* IEEE, 2018, pp. 647–654. doi:10.1109/ICTAI.2018.00103.

[19] D. Toddenroth, T. Ganslandt, I. Castellanos et al., Employing heat maps to mine associations in structured routine care data. Artif Intell Med **60** (2014), 79–88. doi:10.1016/j.artmed.2013.12.003.

[20] W.O. Hackl and T. Ganslandt, Clinical Information Systems as the Backbone of a Complex Information Logistics Process: Findings from the Clinical Information Systems Perspective for 2016. Yearb Med Inform **26** (2017), 103–109. doi:10.15265/IY-2017-023.

[21] J. Cremerius, L. Pufahl, F. Klessascheck et al., Event Log Generation in MIMIC-IV Research Paper. In *Process Mining Workshops,* M. Montali, A. Senderovich, and M. Weidlich, eds. Springer Nature Switzerland, Cham, 2023, pp. 302–314. doi:10.1007/978-3-031-27815-0_22.

[22] L. Auhagen, *Process Mining with routine hospital data: Towards analysis of patient journeys in Oral and Maxillofacial Surgery,* Aachen, 21.12.2023.

[23] A. Berti, S. van Zelst, and D. Schuster, PM4Py: A process mining library for Python. Software Impacts **17** (2023), 100556. doi:10.1016/j.simpa.2023.100556.