Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 196 (2022) 501–508

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021

Developing a Process Mining Tool Based on HL7

João Coutinho-Almeida[a*], Ricardo João Cruz-Correia[b]

[a]*CINTESIS - Centre for Health Technologies and Services Research, University of Porto, Portugal*
[b]*MEDCIDS – Faculty of Medicine of University of Porto, Portugal*

Abstract

In healthcare facilities, processes are not always carried out under the expected methods. The variation in practice leads to lesser quality treatments and greater costs. Within a single visit, patients are now likely to interact with multiple departments, healthcare providers, and Health Information Systems (HIS). Because of these events, information is frequently dispersed, not normalized, and incoherent, creating several barriers to overview processes and audit their quality. Process mining can be a useful tool for getting over some of these obstacles. We designed a procedure to automatically apply process mining techniques using the HL7 standard, which is used for exchanging information between HIS, as a source of event logs. Our work provides a way for pooling HL7 messages from a unified repository of a healthcare institution and provides a pipeline to apply process mining methods to create insights relative to the healthcare processes that are implemented. We show a few diagrams to demonstrate the tool's potential as a process formalization and analysis tool. We concluded that using HL7 messages as a proxy for processes that involve several HIS is a way to easily provide process mining capabilities to an organization.

Keywords: process-mining; healthcare; data mining; HL7; event log

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: joaofilipe90@gmail.com

1877-0509
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the CENTERIS –International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021
10.1016/j.procs.2021.12.042

1. Introduction

Process Mining [1] is a quite recent science, originally designed within the process management field. Its foundations are about utilizing automatically generated event logs in information systems for data analytics. Event logs contain time-stamped records of events, but they can also have attached more information regarding the event or be enhanced through several mechanisms. This is a way for gathering and creating information about the way the processes are being conducted within an organization [2,3].

Process mining is divided into three main branches: process discovery, conformance checking, and process enhancing. Automatic process discovery (i) allows process models to be extracted from an event log; conformance checking (ii) allows monitoring deviations by comparing a model with the event log; and process enhancement (iii) allows extending or improving an existing process model using information about the actual process recorded in the event log [4].

In the healthcare setting, processes are very important. The provision of care services is heavily dependent on the standardization of processes to treat disease and maintain health. Being such a heterogeneous practice, not all processes are according to expected practice [5,6]. Since the number of health information systems in the Portuguese healthcare scene is increasing rapidly, the information that is exchanged between these systems could be useful for creating an event log table. These communications are usually done with heavily standardized procedures like Health Level 7 (HL7). However, these events are often exchanged and are not analyzed, being that the information is often kept (when it does happen) but nothing is done with this information.

This paper describes a project that aims to create a tool for retrieving HL7 messages and creating an event log table out of it automatically. It is also our aim to consequently apply process mining algorithms to this table to extract metrics and knowledge to help reduce the cost of services and improve capabilities, reduce patient waiting times, improve resources productivity, and increase process transparency.

2. Methods

2.1. Event Log Creation

We used some of the HL7 events collected in a Portuguese hospital database in this study. The messages were exchanged between a cancer screening app and the hospital's main clinical software to refer patients who had positive results. Messages were extracted in their original form and converted into an event table. The HL7 messages were being stored in an Elasticsearch database, and the pooling and necessary modifications were carried out using R and Python. We used Python for pooling, orchestration and initial transformation, R for descriptive statistics and BupaR package [7] for the process mining techniques. The algorithms for process discovery were heuristics miner [8] and inductive miner [9].

2.2. Pipeline Development

Apache Airflow was used to create the pipeline. This software is a community-built platform for programmatically authoring, scheduling, and monitoring processes. Workflows are implemented using Directed Acyclic Graphs (DAGs) of jobs in Airflow. The Airflow scheduler executes tasks on an array of workers while following the specified dependencies. The user interface makes it easy to visualize pipelines running in production, monitor progress, and troubleshoot issues when needed. Workflows become more maintainable, versionable, testable, and collaborative when they are defined as code. Figure 1 depicts the total architecture.

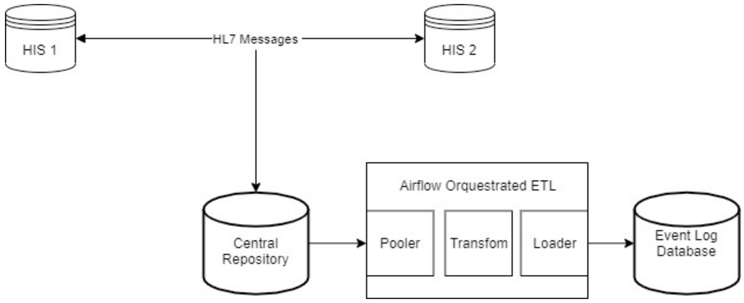

Fig. 1 Architectural Design.

3. Results

3.1. Event Log Creation

The messages collected are all from a laboratory setting. The HL7 messages analyzed were OML (laboratory order trigger event), ORL (laboratory order response event), ORU (Observation Result), and a non-standard message used by the systems in this workflow - ZAK (Report). Relative to the events captured, the messages´ timestamps were analysed. The oldest timestamp is about 2019-02-25 10:35:10 GMT and the most recent is about 2019-12-09 12:56:15 GMT. Regarding message number, the majority are between June and December 2019 and the total number of messages is 390 605. For the creation of events, the fields MSH9 (Message Type), ORC1 (Order Control) and ORC5 (Order Status) of the HL7 messages were used. The correspondence between the fields and events is listed in table 1.

Table 1. Event Mapping.

MSH9	ORC1	Event
ORU^R01	SC	Report Complete
OML^O21	NW	New Order
ORL^O22	OK	Acknowledgment OK
OML^O21	SC	Report Diagnostic
OML^O21	SC	In Progress
OML^O21	SC	Order Complete
OML^O21	SC	Order Canceled
OML^O21	SC	Order Hold
ORL^O22	UA	Acknowledgement NOK
ZAK^ACK	OK	Report Acknowledgement OK
ZAK^ACK	UA	Report Acknowledgement NOK

For the final event log table, additional fields were collected from the messages, namely ORC4 (Placer Group Number) for using as case identifier, ORC10 (Entered by) representing the prescribing physician ID and ORC9 (Date/Time of Transaction) to represent the time of the event. The Lifecycle Transition was fixed at "Complete", and the Activity Instance Id was equal to the row number. The Event Log table created with bupaR was mapped as shown in table 2.

Table 2. Event Log Table.

Event Log Identifier	Original Column	Example
CASE IDENTIFIER	ORC4	4354688
ACTIVITY IDENTIFIER	Event	New Order
RESOURCE IDENTIFIER	ORC10	11925

ACTIVITY INSTANCE IDENTIFIER	Activity Instance Id	5
TIMESTAMP	ORC9	2019-06-01 15:09:39
LIFECYCLE TRANSITION	status	Complete

For the creation of the Event Log table, the cases with a number of events ≤ 2 were removed in an effort to try and reduce unfinished events. For unstarted events, we filtered cases that did not start with the known starting event - "New Order". For the process discovery, the first graph discovered was a dependency matrix shown in figure 2.

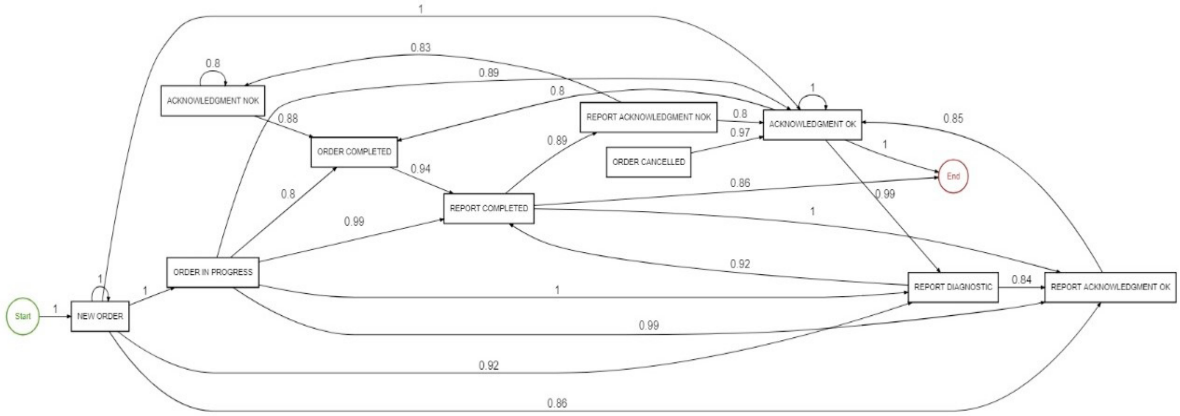

Fig. 2 Dependency Matrix

A precedence matrix is available as well, as seen in figure 3. The most common processes are like the ones seen in figure 4, together they cover 60% of all cases. The remaining images produced with the analysis are all present in the appendixes. Regarding throughput time, where we can see in the **Appendix A** the median time of quartiles of the duration of a case, along with some outliers. The percentage of activity per case is shown in **Appendix B**, which makes it clear that there are some optional steps in the workflow and others that are mandatory.

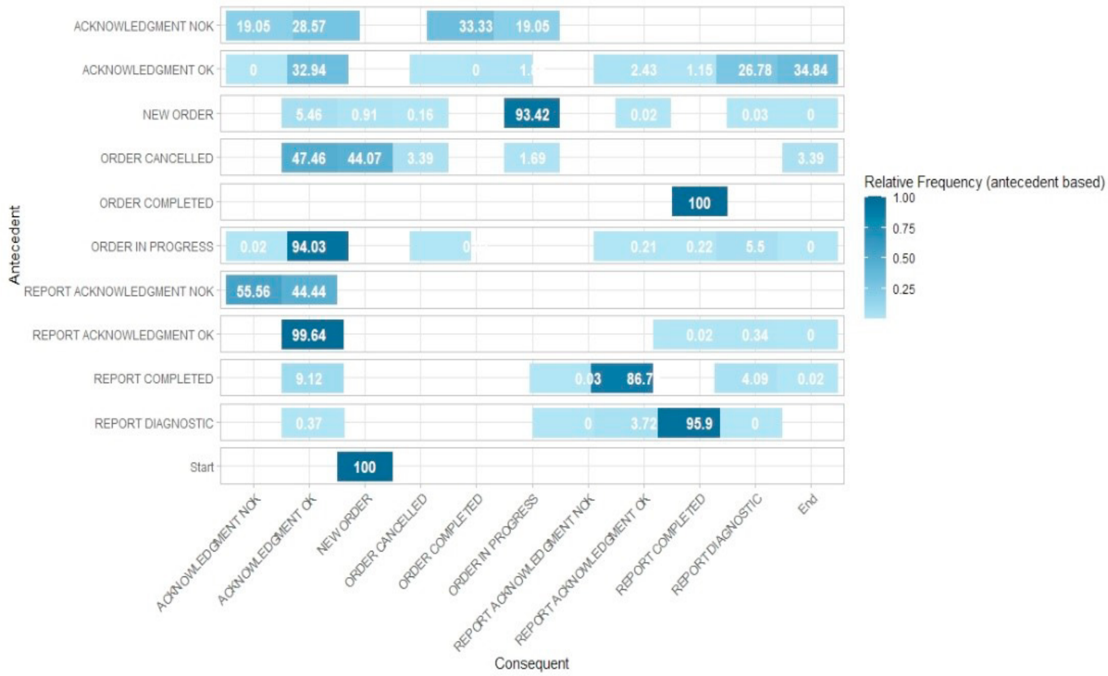

Fig. 3 Relative Frequency Precedence Matrix

Finally, we tried to draw all possible interactions between statuses. **Appendix C** shows the relationship between them with absolute and relative cases present for each node.

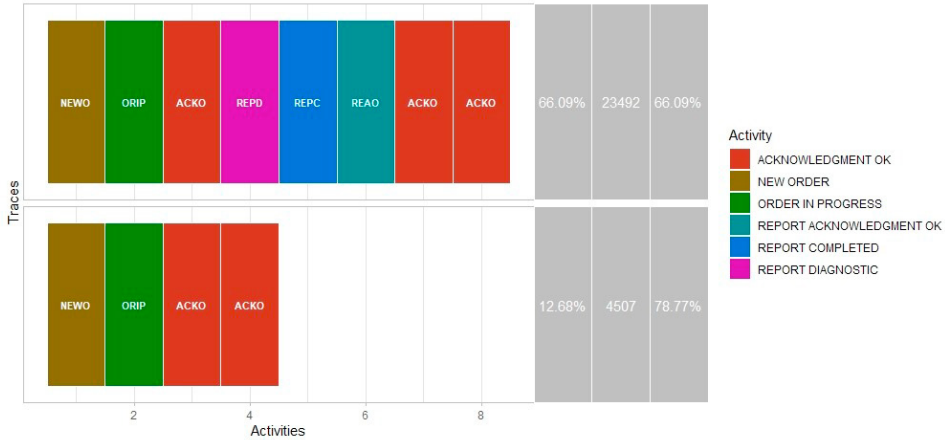

Fig. 4 Activity Relative Frequency Precedence Matrix: First column - relative percentage; second - absolute cases; third - cumulative percentage

3.2. Pipeline Development

The pipeline was created with 3 steps mimicking an ETL process:

- **Extract** Elasticsearch data, receiving as input an Elasticsearch query.
- **Transformation** with accessing elastic documents as JSON, parsing, and creating a new object.
- **Load** with output parameters of the object – csv, database, or cloud providers.

The mechanisms of control flow are included in airflow, processing the DAG every predetermined period. The implementation was made in a development environment, and it is now possible to be implemented in a production environment.

4. Discussion

With this work, an ETL process is possible to be implemented everywhere if an HL7 messages repository exists. The mechanism is robust enough to create the event log table automatically and apply the process mining algorithms in a batch-like manner. Nevertheless, there are some issues, the first and foremost is the data quality. There are messages that have important fields missing. Secondly, the data quantity is also important. At least a year of data would be desirable, for trying to identify seasonality and abnormalities over the course of a year, especially in a case such as the example where a full case takes a very long time (screening). Thirdly, event classification is also an issue to be addressed since the ability to classify cases as completed is very important. However, it is not as simple as classifying as complete all cases with “end events”, since that can restrain outliers that should be analysed. During this work, it became clear that a close relationship with healthcare professionals should be promoted when starting the process mining analysis. Such synergy will be important to tackle classification issues and enlighten cases that may seem odd but can be addressed by someone that works in the institution. Lastly, there is an issue regarding the cut-off period. We do not want to cut the information at a point in time blindly since this will originate a lot of cases that are halfway, whether without start or without end. Even with the removal of all processes with two or fewer events applied along this work, there were still abnormal cases that seem to stem from this data cut-off. Selecting cases with a defined start-event during a “buffer period”, before the actual collection period could be interesting. This would enable to mark cases with starting event and then should be easy to follow up on those for the selected time frame. Cut-off of low-event processes can wield good results, but with a risk of eliminating important information. Finally, the developed system can create an association with different kinds of resources and not only healthcare professionals, so it could be useful to enhance data with that information to create plots and metrics with service providers and/or rooms in order to assess discrepancies as well.

5. Conclusion

With this work, we could understand that HL7 and process mining are a great match to better understand processes in the healthcare setting. Health institutions with several Health Information Systems and thousands of messages flowing daily could benefit greatly from process mining approaches. Even though the example provided was for a single workflow, it can be easily replicated to others, only being necessary a method for mapping messages' content to the event log table. With this, processes can be better monitored, audited, and enhanced, especially if grounded on a good framework for data collection and transformation to provide adequate data for applying process mining tools. This project laid the foundations of said framework to help apply process mining algorithms to HL7 driven event Logs.

Acknowledgements

This work has been done under the scope of the PhD Program in Health Data Science of the Faculty of Medicine of the University of Porto, Portugal – heads.med.up.pt and POCI-05-5762-FSE-000235 - Audit+ project.

Appendix A. Throughput Time

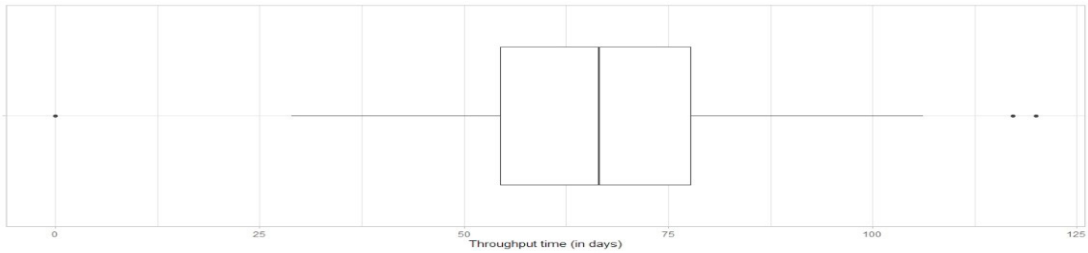

Fig. 5 Case Throughput Time

Appendix B. Case Activity frequency

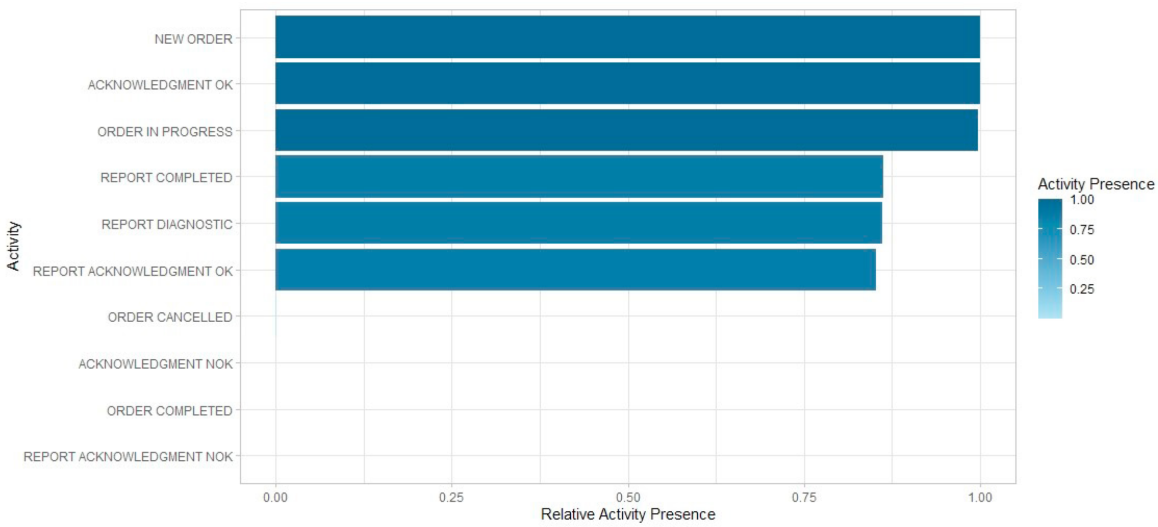

Fig. 6 Activity frequency

Appendix C. Process Mapping

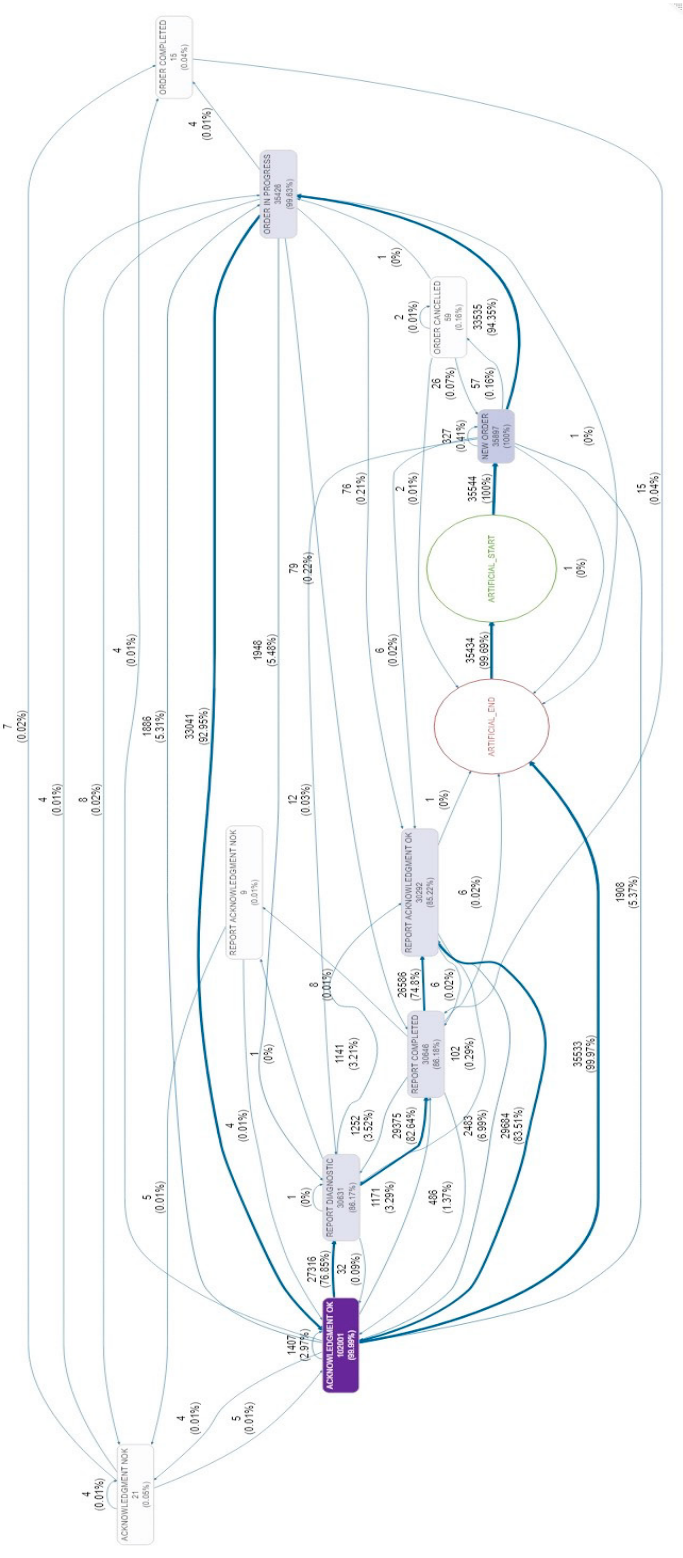

Fig. 7. Process Map: Numbers on top - Absolute cases; in parenthesis - relative percentage.

References

[1] Aalst W. Process Design by Discovery: Harvesting Workflow Knowledge from Ad-hoc Executions 2000. https://www.semanticscholar.org/paper/Process-Design-by-Discovery%3A-Harvesting-Workflow-Aalst/9fb2947768c3b3cc7adef18bf504bf8d1054f2ff (accessed July 29, 2021).
[2] Garcia C dos S, Meincheim A, Faria Junior ER, Dallagassa MR, Sato DMV, Carvalho DR, et al. Process mining techniques and applications – A systematic mapping study. Expert Syst Appl 2019;133:260–95. https://doi.org/10.1016/j.eswa.2019.05.003.
[3] Williams R, Rojas E, Peek N, Johnson OA. Process mining in primary care: A literature review. Stud Health Technol Inform 2018;247:376–80. https://doi.org/10.3233/978-1-61499-852-5-376.
[4] Aalst W van der. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Berlin Heidelberg: Springer-Verlag; 2011. https://doi.org/10.1007/978-3-642-19345-3.
[5] Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D. Process mining in healthcare: A literature review. J Biomed Inform 2016;61:224–36. https://doi.org/10.1016/j.jbi.2016.04.007.
[6] Ghasemi M, Amyot D. Process mining in healthcare: A systematised literature review. Int J Electron Healthc 2016;9:60–88. https://doi.org/10.1504/IJEH.2016.078745.
[7] Janssenswillen G, Depaire B, Swennen M, Jans M, Vanhoof K. bupaR: Enabling reproducible business process analysis. Knowl-Based Syst 2019;163:927–30. https://doi.org/10/gfgvgw.
[8] Weijters AJMM, van der Aalst WMP, de Medeiros; AKA. Process Mining with the HeuristicsMiner Algorithm. Beta Work Pap 2006;166.
[9] Leemans SJJ, Fahland D, van der Aalst WMP. Discovering Block-Structured Process Models from Event Logs - A Constructive Approach. In: Colom J-M, Desel J, editors. Appl. Theory Petri Nets Concurr., Berlin, Heidelberg: Springer Berlin Heidelberg; 2013, p. 311–29.