OJPHI

Process Mining of Incoming Patients with Sepsis

Renee M. Hendricks[1*]

[1]Binghamton University, PhD Candidate, Industrial and Systems Engineering, NY

Abstract

Data mining is a technique for analyzing large amounts of data, in various formats, often called Big Data, in order to gain knowledge about it. The healthcare industry is the next Big Data area of interest as its large variability in patients, their health status and their records which can include image scans, graphical test results, and hand-written physician notes, has been untapped for analysis. In addition to data mining, there is a newer analysis method called process mining. Process mining is similar to data mining in that large data files are reviewed and analyzed, but in this case, event logs specific to a particular process or series of processes, are analyzed. Process mining allows one to understand the initial baseline, determine any bottlenecks or resource constraints, and evaluate a recently implemented change. Process mining was conducted on a hospital event log of patients entering the emergency room with sepsis, to better understand this newer analysis method, to highlight the information discovered, and to determine its role with data mining. Not only did the analysis of the event logs provide process mapping and process analysis, but it also highlighted areas in the clinical operations in need of further investigation, including a possible relationship with patient re-admission and their release method. In addition, the data mining method of creating a histogram, of the process data, was applied, allowing data mining and process mining to be utilized complimentary.

Keywords: Data mining, process mining, sepsis

*Correspondence: rhendri2@binghamton.edu

DOI: 10.5210/ojphi.v11i2.10151

Introduction

In the healthcare industry, as in other industries, there are vast amounts of data collected and stored, but left unanalyzed. This is ill-timed as the data may provide historical information or trends that could assist with developing current and future methods, strategies and prediction models. This is where data mining assists as trends and patterns can be understood from these large amounts of data, as well as providing ways to classify or identify similar groups or entities. In healthcare, patients can be identified or classified based on a series of health measurements to determine who is at risk for developing a particular disease, or at what disease stage they are approaching. Data mining is effective but it's limitation may be in mining process or time-stamped datasets.

When analyzing event logs or other time-stamped data, the entities or steps are dependent and build upon other information in the same file, so this requires the event logs, such as for one patient or one medical staff, to be analyzed as a whole unit and not separated out as typically conducted with patient files for analyzing trends or prediction in disease. Granted, in data mining, there can be dependent and independent variables, but the information in event logs are not well defined nor bounded variables, but a series of steps that could be labeled as variables but separating them from their pre- and post- steps provide no meaning and therefore requires analysis as a whole. A newer method of data analysis, called process mining, has come of interest for analyzing event log, data formats.

Process mining is conducted to gain knowledge about a particular process or series of processes being executed. Having an accurate model of the process behavior improves the implementation and evaluation of the process as well as configuring any additional requirements not included in the system [1]. Based on a Dutch hospital dataset available online, this project applied process mining methods to this event log, to first understand this method and outcomes, but to also determine its relationship with data mining. Is a separate mining method required for process data or can data mining methods also be utilized? Are process mining and data mining complimentary? The process data for this study was a record of the patients' steps when entering the emergency room (ER) with sepsis cases, as well as their release and possible re-admission steps. Process mining is a new, data method area in need of applied examples in order to understand how to use it, the expected outcomes, how to improve it, and its role with data mining.

Process Mining

When searching the topic of data mining in healthcare, there were a series of articles that referenced process mining. Many of the articles were conference proceedings, which may indicate the newness of this method. Process mining is defined as a method for exacting data from information systems, such as hospital information systems, to gain understanding about the processes and further refine them. Healthcare processes are complex and involve steps executed by personnel from various disciplines and offices. This complexity makes it interesting yet difficult to analyze and understand. Process mining is defined as extracting knowledge from data generated and stored in information systems in order to analyze them [1]. Typically, event logs are utilized for analysis, as they provide timestamps of steps as well as the personnel or system users who completed the steps, along with any step completion information, such as a mouse click, data entered, or an item selected. Because of the use of recorded event logs, process mining is considered an a-posterior analysis or analysis after the event(s) [2].

The authors in [1] found a plethora of information on process mining. Both types of processes in healthcare can be analyzed with this method: medical (clinical) processes and organizational (administration) processes. The software tools typically utilized in process mining are ProM and Disco, both which can be downloaded for free. In addition, three common algorithms that are utilized in process mining are fuzzy minor, heuristics minor and trace clustering. Heuristics and fuzzy minor are process discovery methods whereas trace clustering in similar to clustering in data mining, in that events or steps are clustered.

The authors in [1] also list many benefits of conducting process mining, especially in healthcare. Process mining can assist with understanding and even predicting both the staff and patient behaviors in certain situations as well as assist in redesigning and improving the process. Process mining provides information on what is causing the bottlenecks and allows one to analyze process performance and reduce times, such as patient wait times and procedure times. Most importantly, process mining can determine the gap of what is supposed to happen in a process(es) versus what is actually happening in the process(es), so the process(es) can be better understood and improved.

Process Mining Case Studies

One of the case studies found regarding process mining determined the diagnostic, therapeutic and clinical processes from hospital admission to discharge of 368 patients who were diagnosed with a first-ever stroke at four hospitals in Italy [2]. The event logs from the hospital chart system were utilized and the heuristics minor algorithm was chosen to map the process steps taken by the patients. The timeframe of this study was not provided, so the length, in weeks or months was unknown, but the study still provided valuable information as the process maps for each of the 4 hospitals were determined. This provides not only the process steps, for each hospital, but the entire, average process time from admission to discharge for the patients, along with the standard deviations among patients. This study also allowed the hospitals' processes to be compared to understand the different steps taken by each, something that is new in the healthcare industry. Further analysis of these steps may provide additional information into the clinical practices and how they can be refined. Typically, this information is reviewed and utilized from a clinical operations perspective, to make sure correct protocols are followed, but this same information is not typically utilized for process understanding and improvement, making process mining a new research area and possibly an effective new method [2].

A second case study of process mining determined the common paths taken by patients arriving for an outpatient procedure. The researchers' intent was to first determine the most common process and steps and then compare to an expert created process. The event logs were taken over a month time-period from the hospital information system. The researchers were able to extract 699,136 event logs and 123,299 patient cases [3]. In this case, both the heuristics and fuzzy minor algorithms were utilized as the researchers wanted a more refined model of the process from the use of 2 techniques, along with the use of both software programs, ProM and Disco. The researchers discovered in this case that the most common path taken by outpatients was: consultation registration, consultation, consultation scheduling, payment, and then outside hospital prescription printing. The researchers also found that the process developed with the algorithms matched the expert created model by 89%, a significant result [3].

The intent of process mining falls in line with process improvement and lean thinking in that the process needs to be understood, controlled and evaluated for continuous improvement. Process mapping and lean thinking has been utilized for decades to sketch the process or processes under study, in order to determine the process steps and functions over time to analyze a situation, but very few individuals have reviewed event logs to see if more information can be gathered about the process. Attempting to sketch a process map from a large event log, even for a week of events, would be difficult and time consuming for an individual or team, but may be efficiently completed with process mining software tools. As in lean thinking, if the process is already understood and

documented, then process mining may allow others to improve the process or understand the effects of a change made prior to implementation. Process mining and process mapping are not seen as replacements for each other, but as compliments as both techniques provide information that can highlight and improve the process. In addition, event logs may depict a different picture that cannot be gathered from observations or interviews with participants, providing a reliable, constant source of information.

Methods

The purpose of this study was to analyze a hospital's event log of patients entering the emergency room with sepsis, using a process mining software tool and algorithms, to understand how process mining is conducted and to interpret the results. In addition, the researcher wanted to determine if data mining methods can assist with analysis of event logs.

Data Description and Format

The data utilized for this study was from a Dutch hospital event log downloaded from the 4TU.Center for Research Data website [4]. This information is from the hospital's enterprise resource planning (ERP) system and includes 15,214 events, for 1,050 patient sepsis cases, from November 7, 2013 to June 5, 2015 [5]. In addition, the events are divided into 16 hospital activities, or classes, with one case representing one patient's pathway through the hospital [5].

The downloaded file was compressed and of the .xes extension, which is an event log file format. Because existing zip applications do not open this file, the z7-zip application was downloaded in order to extract the event log file (.xes) from the compressed file. When searching for studies that utilized this same dataset, it was discovered a few studies utilized this same data, but for only determining and mapping the process flows. Data mining methods were not utilized in the previous studies but are utilized in this study.

Tool, Initial Data Review, and Algorithms

ProM Tool

The source tool utilized for this study was ProM 6.7, revision 35885. ProM is the abbreviation for Process Mining Framework [6] and was downloaded from the process mining website [6]. ProM was selected for this study as it was noted in [1], that ProM is the most commonly utilized tool in healthcare. There are two versions available, a lite version, ProM Lite 1.1 and the ProM 6.6 and 6.7. Most times, a lite software version has limited usage and tools, so the 6.7 version was selected. In addition, the 6.6 and 6.7 versions are targeted for researcher use. ProM also has a package manager tool, where many of the algorithm packages are selected and downloaded, so they can be utilized. This is similar in programming, when listing the libraries in a command line, so the functions and algorithms can be executed in the program. In addition, the ProM tool can accept various event log file formats, including the .xes format, and also allowed the .xes file to be saved as a .csv file, which can be viewed and analyzed using tools. MATLAB is a current data mining and analysis tool that was also utilized in this study for a quick overview of the data in the event log.

Initial Observations of the Event Log Data

First, the sepsis event log file was imported into ProM 6.7. Before any analysis, the file was also exported as a .csv, for later use. ProM provides a series of functions based on the data imported. Available algorithms are highlighted green in a function listing. Functions can also be found through a search field at the top. To understand the overall data, the View Source option was selected and a dashboard of the data was provided, as seen in Figure 1 below. One process is listed, with 1,050 (patient) cases, and 15,214 events, along with 16 classes (of events). This confirms that the data is as expected and matches the data description in [5].

In addition, the data time frame from November 7, 2013, to June 5, 2015 is also listed correctly. This dashboard also shows that the events per overall cases range from a minimum of 3 events (or steps), a mean of 14 events and a maximum of 185 events. The large number of events per case may be associated with the patients who were re-admitted as a multitude of events would have occurred. In addition, the events are divided into 16 event classes or activities, with a min of 3 event classes per patient case, a mean of 9 event classes or steps per patient case and a max of 12 steps (of the total 16) per patient case. This makes sense as not every event class (step) will be followed by any one patient.

Figure 1: Sepsis Event Log Dashboard in ProM

After confirming the data characteristics, the *Summary* option was selected and this provided a listing of the 16 activities or steps for the patients in this dataset. As can be seen in Figure 2 below, the 16 activities consist of some of the following: ER registration, IVs or tests administered, being admitted to the triage, and the series of releases A-E, and if the patient returned to the ER. Also, the 16 steps are listed by their occurrences from high to low, with Leucocytes and CRP having the highest percentages of occurrences at 22.236% (3,383 occurrences) and 22.441% (3,262 occurrences) accordingly, whereas many of the ER release steps occurred in lower frequencies. Leucocytes, Lactic acid and CRP are all measurements provided by blood tests. Leucocytes are white blood cells and a high white blood count can signal an infection the body may be fighting [7]. CRP, or C-reactive protein can signal inflammation in the liver if the test result is high [8]. If these patients had sepsis or another infection, blood tests are typically conducted to see if there is a sign of an infection and it determines the path forward for treatment. In addition, there are two types of admission steps a patient can take depending on their conditions. This is because the patient can be admitted to 2 units, Intensive Care (IC) unit, or not as intensive (NC) unit. There are also five release types, A-E, for the discharge path.

concept:name

Event classes defined by concept:name
All events

Total number of classes: **16**

Class	Occurrences (absolute)	Occurrences (relative)
Leucocytes	3383	22.236%
CRP	3262	21.441%
LacticAcid	1466	9.636%
Admission NC	1182	7.769%
ER Triage	1053	6.921%
ER Registration	1050	6.902%
ER Sepsis Triage	1049	6.895%
IV Antibiotics	823	5.409%
IV Liquid	753	4.949%
Release A	671	4.41%
Return ER	294	1.932%
Admission IC	117	0.769%
Release B	56	0.368%
Release C	25	0.164%
Release D	24	0.158%
Release E	6	0.039%

Figure 2: Sepsis Event Log Summary of Activities in ProM

To test the initial analysis of the event log in MATLAB, the .csv format of the sepsis event log file was imported into MATLAB and the histogram function was selected. The output can be seen in Figure 3, on the next page, and it matches the activities summary in ProM, with the Leucocytes and CRP steps having the highest occurrences and the patient discharges having the lowest frequencies. So, it appears MATLAB may assist with process or event log data, at least for initial

analysis or frequency counts. This also confirmed that the original data file, which was converted to the .csv format, in ProM, did not lose any data.

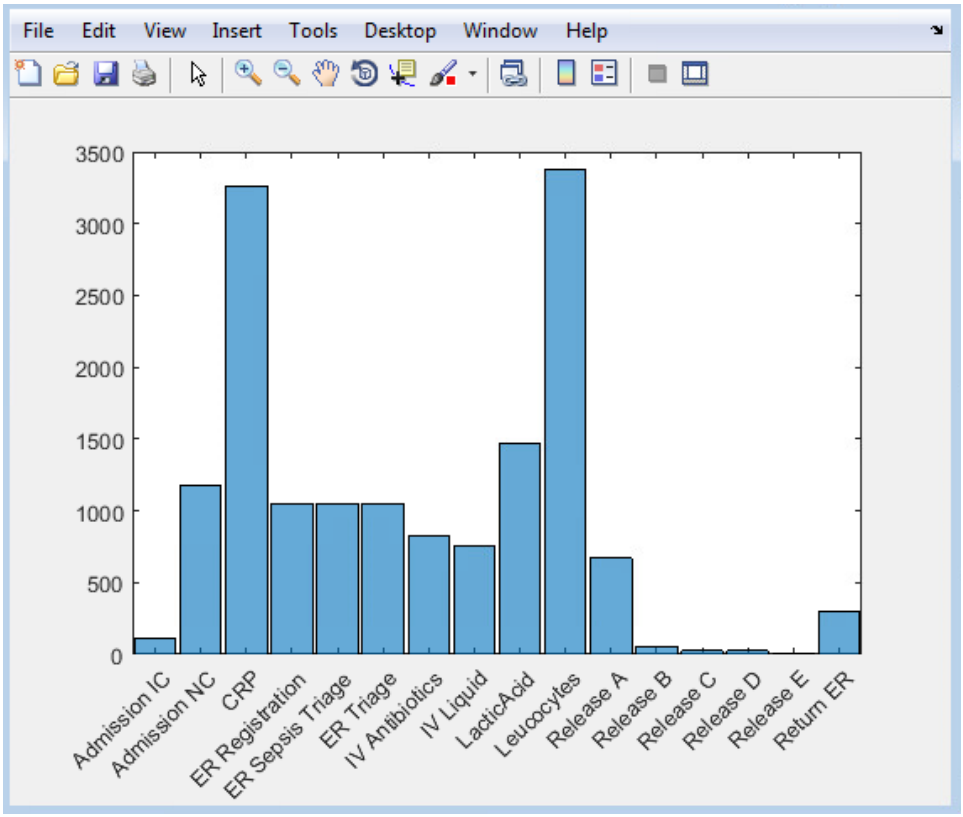


Figure 3: Sepsis Event Log Histogram of Activities in MATLAB

ProM Algorithms

Using the ProM software tool, a series of algorithms, or functions were applied to the sepsis event log and included first determining the process map or model, along with discovering the casual matrix. To create the process map or model, the *Alpha Minor* function was selected. This feature is a discovery approach for mapping the process of the event logs. This is the first time this researcher utilized these event logs, so it is necessary to discover the path first. If past or known results were gathered, then the discovery phase could be skipped. The event logs are the steps the patient encountered while being admitted to the ER and being tested for sepsis, and then possibly released or admitted to a hospital unit. The results of this function are explored in the next Results section. This feature was utilized to visualize the process map of this event log before conducting any further analysis.

To determine any causal relationships of the activities, the *Discover Matrix* function was selected. This calculates a value from 1 to -1 for each of the 16 activities. A 1 (ideal) value means that there is a casual relationship between the row activity and the column activity. A -1 value means there is no causal relationship between the row and column activities.

Results

The process model or sometimes called, Petrinet, for the sepsis event log was created using a discovery algorithm. The results of this mapping can be seen in Figure 4 below. The process starts at 1, which can lead to a series of steps, whether Admission IC, CRP test, as well as the ERP Triages, ER Registration, and IV Liquid, as each patient follows a different route, based on their condition and encounter. The arrows represent the transition or next activity in the process. Also, it appears in some cases, that after the blood test results of CRP and Leucocytes, patients were then released. As expected, there are patient cases where different admissions (IC and NC) were utilized, as well as different discharge types (A-E) and some cases where patients were re-admitted. The circle at the bottom of the figure is the end point of the patient process.

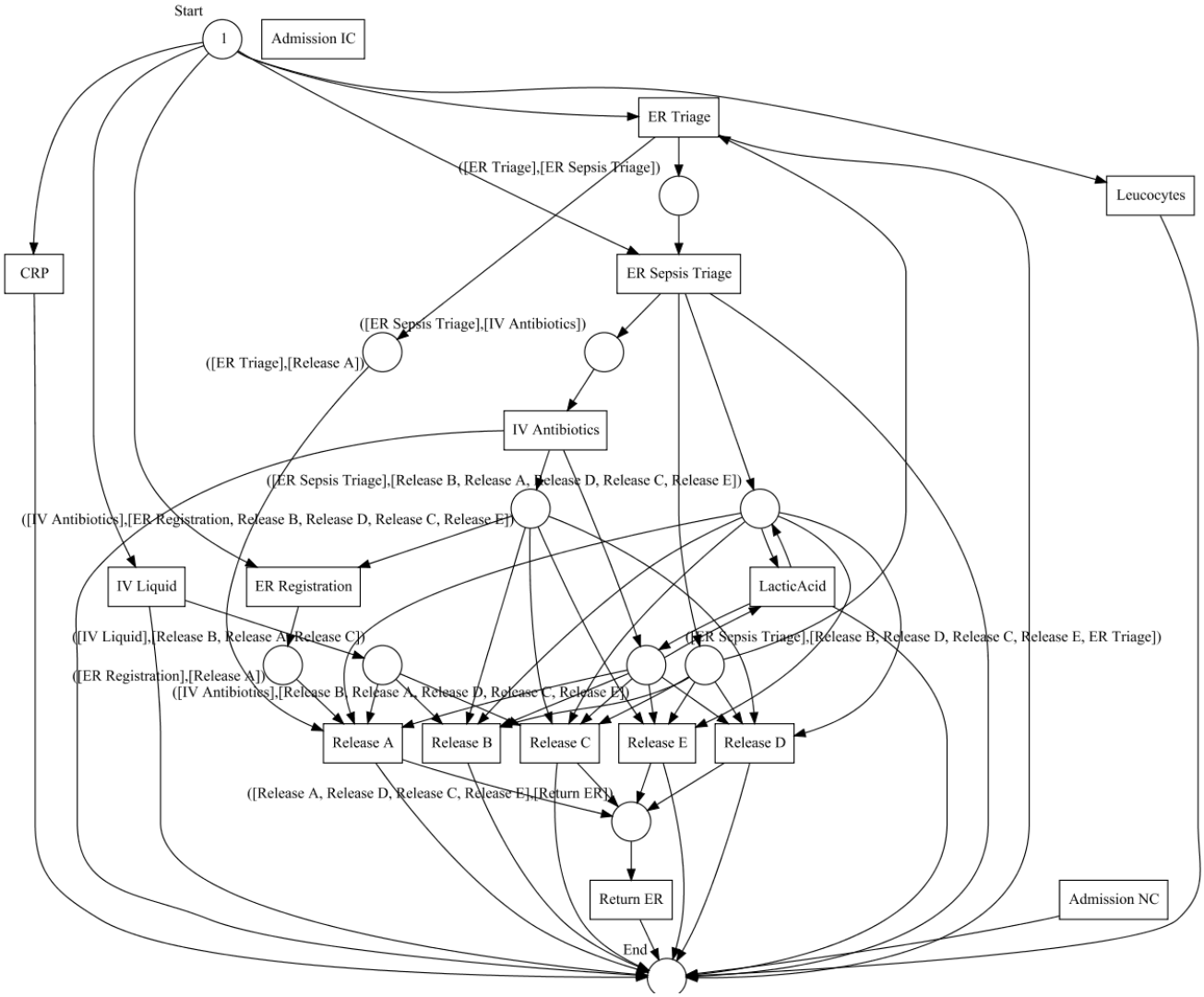
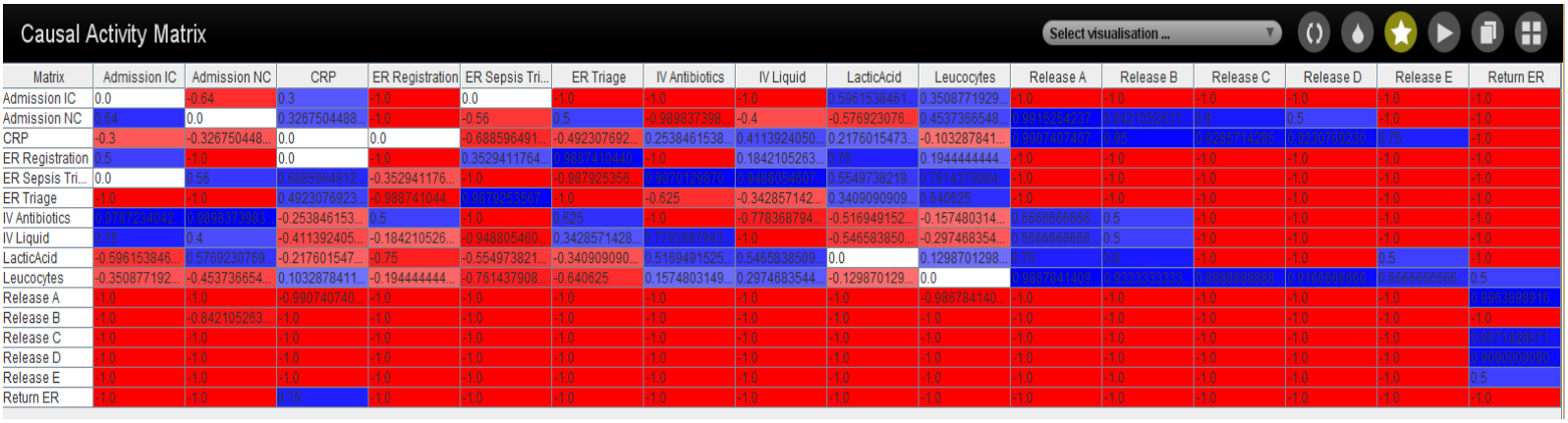


Figure 4: Sepsis Event Log Process Model/Petrinet in ProM

In addition, the *Discover Matrix* function was utilized to see if there are any causal relationships between the 16 activities. A 1 value means there is a causal relationship between the row activity and the column activity. The results of this test can be seen in Table 1 below. The heuristics minor, a discovery method, was utilized. The values higher than 0 are highlighted blue, as there may be a causal relationship. Looking at the rows, the step, Leucocytes has a series of values above 0.9, which means that this step has a causal relationship with a series of columns, which includes all the release steps and the return to ER steps. This makes sense as many of the patients require the blood test to discover if there are any underling infections the body may be fighting, before discharged. Many cases where there is a 0.0 value (in white), this means the step cannot be compared to itself, similar to a correlation matrix where the values are left empty. It is also not surprising that the CRP test also has high causal relationship with the discharge steps as this tests for any possible infections, prior to patient release. It also appears Admission NC, ER Registration, IV Anti-biotics and IV Liquids have a causal relationship with Admission IC, with a value of 0.64, 0.5, 0.98, and 0.75 accordingly. It also appears that some of the releases (A, C, D, E) have a causal relationship with re-admittance into the ER, which may require further analysis. It is important to understand not only the steps but the profiles of the patients that were re-admitted to the ER, so as to reduce re-admission. Reducing the re-admission levels reduces the work load on the medical staff, reduce the patient recovery times, along with reducing the costs incurred by all.

Table 1: Sepsis Event Log Casual Relationship Matrix in ProM



Matrix	Admission IC	Admission NC	CRP	ER Registration	ER Sepsis Tri...	ER Triage	IV Antibiotics	IV Liquid	LacticAcid	Leucocytes	Release A	Release B	Release C	Release D	Release E	Return ER
Admission IC	0.0	-0.64	0.3	-1.0	0.0	-1.0	-1.0	-1.0	0.5981538461...	0.3508771929...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
Admission NC	0.64	0.0	0.3267504488...	-1.0	-0.56	0.5	-0.889837398...	-0.4	-0.576923076...	0.4537366548...	0.9810524237...	0.8421052631...	0.8	0.5	-1.0	-1.0
CRP	-0.3	-0.326750448...	0.0	0.0	-0.688596491...	-0.492307692...	0.2538461538...	0.4113924050...	0.2176015473...	-0.103287841...	0.9907407407...	0.95	0.9285714285...	0.9230769230...	0.75	-1.0
ER Registration	0.5	-1.0	0.0	-1.0	0.3529411764...	0.9867410440...	-1.0	0.1842105263...	0.75	0.1944444444...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
ER Sepsis Tri...	0.0	0.56	0.6885964912...	-0.352941176...	-1.0	-0.987925356...	0.9875129875...	0.9486842607...	0.5549738219...	0.7614379084...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
ER Triage	-1.0	-1.0	0.4923076923...	-0.888741044...	0.8878256607...	-1.0	-0.625	-0.342857142...	0.3409090909...	0.640625	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
IV Antibiotics	0.9767234042...	0.8898373983...	-0.253846153...	0.5	-1.0	0.625	-1.0	-0.778368794...	-0.516949152...	-0.157480314...	0.6666666666...	0.5	-1.0	-1.0	-1.0	-1.0
IV Liquid	0.75	0.4	-0.411392405...	-0.184210526...	-0.948805460...	0.3428571428...	0.7783687843...	-1.0	-0.546583850...	-0.297468354...	0.6666666666...	0.5	-1.0	-1.0	-1.0	-1.0
LacticAcid	-0.596153846...	0.5769230769...	-0.217601547...	-0.75	-0.554973821...	-0.340909090...	0.5169491525...	0.5465838509...	0.0	0.1298701298...	0.75	0.8	-1.0	-1.0	0.5	-1.0
Leucocytes	-0.350877192...	-0.453736654...	0.1032878411...	-0.194444444...	-0.761437908...	-0.640625	0.1574803149...	0.2974683544...	-0.129870129...	0.0	0.9867841409...	0.9811320754...	0.9696969696...	0.9166666666...	0.6666666666...	0.5
Release A	-1.0	-1.0	-0.990740740...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-0.986784140...	-1.0	-1.0	-1.0	-1.0	-1.0	0.9963099630...
Release B	-1.0	-0.842105263...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
Release C	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	0.9857142857...
Release D	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	0.9969969969...
Release E	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	0.5
Return ER	-1.0	-1.0	0.75	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0

Conclusion

In conclusion, process mining methods and algorithms appear effective in analyzing the available hospital event log. Not only does analysis of the event logs provide process mapping and process analysis, but also highlights areas in the clinical operations that may require further investigation. Modeling the patient path helps one to understand the steps the person followed, and the steps the medical staff followed in their roles, to both learn the process or processes under investigation as well as to refine the process with the intention to reduce patient waiting times and re-admissions as well as streamline the clinical processes for the medical staff. In addition, it also appears that some data mining methods, such as creating a histogram, is also of use with process data. So, data

mining methods and process mining may be utilized complimentary in future event log analysis. But process mining may be a division in its own right, as the event log data requires different analysis than independent variables, such as patients results, require in data mining. So, it appears, both data mining and process mining will be here to stay, for now.

The limitation of this study is that one dataset, which is one hospital event log, was utilized. Also, two software programs, ProM and MATLAB, were utilized for this study. It is recommended for future research that multiple events logs are analyzed, using both process mining and data mining methods, whether for the same process, so comparisons are made for a particular process, or across processes, units or hospitals, to compare the software results, to better understand how the algorithms work with these data formats, as well as to gain insights to these processes under investigation, in order to baseline and improve them. With limited software options available when analyzing event logs, it is difficult to accurately confirm the results, repeatedly.

Acknowledgements

I would like to thank my course Professor Dr. Won, and my Department Head, Dr. Khasawneh because their insights provided me an interest in data and process mining and machine learning.

References

1. Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D. 2016. Process mining in healthcare: A literature review. *J Biomed Inform*. 61, 224-36. PubMed https://doi.org/10.1016/j.jbi.2016.04.007

2. Mans R, et al. 2008. Process mining techniques: an application to stroke care. *Stud Health Technol Inform*. 136, 573-578. PubMed

3. Kim E, Kim S, Song M, Kim S, Yoo D, Hwang H, Yoo S. 2013. Discovery of outpatient care process of a tertiary university hospital using process mining. *Healthc Inform Res*. 19(1), 42-49. PubMed https://doi.org/10.4258/hir.2013.19.1.42

4. 4TU.Centre for Research Data. 2017. http://data.4tu.nl/repository/. Accessed 30 November 2017.

5. Mannhardt F. (2016) Sepsis Cases - Event Log. Eindhoven University of Technology. Dataset. https://data.4tu.nl/repository/uuid:915d2bfb-7e84-49ad-a286-dc35f063a460

6. Eindhoven University of Technology. 2017. http://www.processmining.org/start Accessed 15 November 2017

7. MDHealth.com. 2017. http://md-health.com/leukocytes.html Accessed 10 December 2017.

8. Web MD. 2017. https://www.webmd.com/a-to-z-guides/c-reactive-protein-test#1 Accessed 10 December 2017