Process Mining using bupaR package using Healthcare Processes data

SNEHAL AGRAWAL

Carl H. Linder College of Business, University of Cincinnati, Cincinnati, USA

#### Abstract

Process mining is a part of data science that allows organizations to discover, analyze and improve their business processes. It focuses on the analysis of event data produced during the execution of processes in any business. Hence, process mining provides the link between data mining or machine learning techniques and the business process management system. The techniques, algorithms and software packages intended for the application of this discipline have been emerging undoubtedly over the past two decades. However, most tools cannot create workflows that can be reused later to replicate the results, and most tools are not easily customizable. This paper introduces bupaR which is a collection of R packages that provides a platform for reproducible process analysis in R and supports various phases of a process analysis project, ranging from data extraction to data analysis. It helps in exploring and visualizing event data and monitoring processes. Furthermore, it stimulates the application of process mining in organizations, as R is a very widely used analysis tool in business, with its popularity still strongly increasing.

**Keywords:** Process Mining; Data Science; Data Analysis; R; Event log; bupaR; Healthcare

### 1 **Introduction**

Process mining has gained a lot of attention from both academics and industry, which led to the growth of a series of data mining algorithms and techniques to allow analyzing business processes. This field of data science aims to provide a broad summary of processes in a short time, explain unstructured and multidisciplinary processes, design advancement opportunities for processes, and incorporate domain knowledge into process validation. Process mining has led to the development of several commercial and open-source process mining tools to support the analysis of the actual event logs that are available in information systems. The majority of these tools comprises of the discovery of a process model that precisely defines the process as captured within the analyzed event data, conformance checking to what extent a given process model is perfectly describing event data, and enhancement procedures for process models by projecting engaging information on top of a model.

This paper focuses on process analysis to help extract the processes involved in an event and make use of tools to redesign the system as the businesses want it 'to-be'. However, the process mining techniques provide important contributions to the preparation and cleansing of the event log. When an enterprise resource planning system is applied to the inventory management process, the system records all the activities involving the organization of the warehouse, purchase order and approval, storage and delivery of the goods in the inventory along with the starting and finishing times recorded by the people who performed each activity. In this way, process flows can be reproduced based on their actual execution. This information is obtained from the event logs of the system and it can be assessed by various data analysis techniques so that businesses can identify, create, record, and enhance their processes. Therefore, the ultimate purpose of information systems and data science is not just to gather more data, but to turn data into real value. The aforementioned suggests that data should be utilized to create new or enhance existing products, services, and processes. The final target of process mining is the improvement of business processes, mainly by applying analytical tools.

Over the past few decades, the open-source analytical language R has seen a massive rise in popularity not just among researchers in the field of data sciences but also within businesses. The gained popularity of R is mainly because of the provision of a large number of packages that extend the functionalities of the codes and save time. One of the packages in R is bupaR (Business Process Analysis in R) used for process mining. bupaR is an open-source package in R for handling and analysis of business process data. It was developed by the Business Informatics research group at Hasselt University, Belgium. The package includes basic functionality for generating event log objects and to get information about an event log. bupaR aims at assisting at each step in the analysis of event data, from data acquisition to process monitoring and process modeling. To perform the process analysis, the bupaR package uses some of the related packages in R such as eventdataR, xesreadR, edeaR, processmapR, and processmonitR, each of these packages has a specific purpose to contribute in the process mining.

Process mining applies computer science and mathematics to analyze processes based on data carried in event logs. Process mining is aimed at answering questions about operational processes in various domains including healthcare. For example, what happened in the past and why, what will happen in the future, how can a company better control its processes, and how can a process be redesigned to improve performance. In healthcare process management, healthcare experts intend to formalize, develop, and automate processes. Given that in information management systems nearly every phase in clinical procedures is performed digitally and documented (as event logs), healthcare data may provide insights into the system's vulnerabilities to recommend changes. Therefore, there is an urge to show that process mining can be used to reduce healthcare's operational costs. We use patients' data from the eventdataR package of bupaR to demonstrate the business value of process mining. The log specific to the healthcare process is gathered and analyzed. Using the R libraries, we derive a process model based on the data. The resulting process model is then used to offer suggestions based on variations from the expected business model and obstruction points found in the analysis, which illustrate process mining's market value.

### **2** **Available process analysis tools**

We will address some of the tools used during process analysis and mining. ProM and Apromore both provide an easy to use graphical user interface for non-expert usability. Though the user-friendly interface attracts the non-expert users to use and showcase process mining capabilities to a broader audience, it hinders the usability of the tools for large-scale scientific experimentation. To this end, RapidMiner, which allows for the definition and execution of analysis workflows, is connected to the process mining framework ProM that formed the RapidProM tool to allow for repeated execution of large-scale experiments with process mining algorithms. However, RapidProM does not provide a way to either customize an algorithm or integrate custom-developed algorithms. As such, the stated tools fail to support customizable process mining algorithms and large-scale experimentation and analysis [1]. We propose the framework for Process Mining in R (bupaR) to address the lack of previously mentioned resources as bupaR is easily extendable and allows for algorithmic flexibility and large-scale experiments. The bupaR library supports process mining in the statistical language R, which is widely used in data science.

### **3 Concepts of process mining**

The following sections present the building blocks of process mining. We start by explaining an event log, and then we define the process model and its benefits. This information gives the reader the fundamental details needed to understand our project.

#### 3.1 Event Log

The key data source within process mining is event data, which is also referred to as an event log. The starting point for process mining is an event log [2]. The event log is the historical record of the execution of the business processes, containing all their different instances like Order 121, the activities involved like receiving the goods from the supplier, the people in charge of each activity, and the starting and finishing times of each activity and other associated data like the product to be purchased, supplier or requesting department [3].

Event logs capture the following fundamental information along with other details:

###### 3.1.1 Activity

Activity is a well-defined step in a process. For example, in hospital admission, the activities would include registration, blood test, discharged.

###### 3.1.2 Case identifier

A case identifier is a unique identifier that allows activities to be tagged to the respective case. For example, in hospital admission, the case identifier will be the patient identifier.

###### 3.1.3 Activity instance identifier

The activity instance identifier is a unique identifier that allows an activity instance to be tagged to the related activity. The activity instance identifier allows activity instances within a case to be arranged in sequential order. The activity instance is more granular than activity. There can be more activity instances than activities for a specific case. For example, a patient has an activity “X-ray” which is 1 activity, but it can have 2 activity instances. The first activity instance is when the patient’s X-ray started, and the second activity instance is when the patient’s X-ray ended.

#### 3.2 Process Model

Process models describe the ordering of activities in a process [2], likely augmented by additional details such as time limits, use of resources, or use of data. Through analyzing the event logs of activities, we discover a process model. A process model integrates all the instances in an event log to reflect the event log's common actions as a single operation. Modeling business processes is the graphical representation of business processes or workflows in an organization, as a means of identifying possible improvements.

#### 3.3 Method

The method of process mining involves the following steps: (1) defining the scope and goals of the project and questions to be answered, (2) extracting the data on cases, activities, and timeframe, (3) formatting event data as event logs and abstracting or filtering as sub-processes, (4) visualizing event logs in a log/pattern investigation activity to obtain the limitations and opportunities of the datasets, (5) applying the process mining techniques to find answers, (6) evaluating the answers and gathering the scope of improvements, (7) using the findings to create process improvement proposals [4].

### **4** **Software architecture and functionalities**

bupaR consists of various R packages, each with a specific purpose to fulfill in the process analysis. The package bupaR itself acts as the core of the suite, providing the primary features used by the other packages for handling event data. The packages used in the process mining in R are listed below and discussed briefly in the next section.

#### 4.1 bupaR

bupaR provides support for various stages in process analysis, such as importing event data, calculating descriptive, process monitoring and process visualization. bupaR includes basic functionality for generating event log objects in R. An event log in this context is a dataset combined with a mapping. The mapping contains the properties of the specific event log, i.e., the case identifier, activity identifier, timestamp, etc. Each element of this mapping refers to a variable in the data set.

#### 4.2 edeaR

edeaR stands for Exploratory and Descriptive Event-Based Data Analysis, and its purpose is to perform more in-depth analyses of event logs. It basically contains two sets of functions: process metrics to describe and explore event logs and data filters for data subsetting. The process metrics are based on Lean Six Sigma literature and can be analyzed and visualized at different levels of granularity [5].

#### 4.3 eventdataR

The eventdataR package provides a data repository with both artificial and real-life event logs to experiment with new techniques and algorithms. Currently, both artificial event data such as patients dataset, as well as real-life event data, such as the sepsis dataset, exist in this package. Each of them can be loaded very easily with the data function in R [5].

#### 4.4 xesreadR

The xesreadR package allows to read and write xes-files. This makes bupaR compatible with the IEEE standard for sharing and storing event data, and thus with other process mining tools. For example, this is beneficial if we want to combine bupaR with RapidProM, as RapidMiner also supports the use of R-scripts [5].

#### 4.5 processmapR

The package processmapR provides a means of creating visualizations of the process, such as process maps and dotted charts, using customizable map profiles [6]. The visualizations produced are highly customizable and can be used to provide insights into various aspects of the process. The default profile is the frequency profile on both arcs and nodes, with absolute frequencies. The combination of processmapR functions with edeaR subsetting functions will simplify complex graphs.

#### 4.6 processmonitR

processmonitR provides a limited set of process dashboards. These can be used in a permanent, real-time fashion, as well as for interactive data analysis. The available dashboards, which build upon the Shiny framework [7], are largely organized along with the metrics of edeaR.

### **5 Applications and illustrative examples**

There are vast volumes of data collected and stored in the healthcare sector, as in many sectors, but left unanalyzed. This is untimely as the data may provide historical details or patterns that could assist in the development of current and future models of processes, strategies, and predictions. Healthcare processes are complex and involve actions carried out by workers from multiple backgrounds and offices. This complexity makes examining and understanding the dynamic, complex, and cross-functional processes fascinating but difficult.

In this paper, we are examining the event log embracing the events of patients from a hospital. The dataset contains the events from when a patient enters the hospital until he/she exits. The objective is to investigate this healthcare process for two stances: control flow and conformance examination. Process mining techniques are used for the exploration and summarization of the patients with the event log. We have extracted the useful facts from event logs and scrutinized the pathways followed by distinct processes. The results of these stances provide new insight that facilitates the refinement of the current process.

#### 5.1 Prerequisites

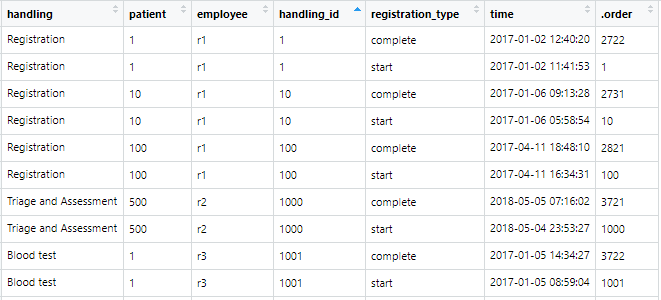
In order to use bupaR, R and Rstudio (or another IDE) need to be installed. Then installation of bupaR can be done easily by executing install.packages("bupaR") in the R-console and then loading the library(bupaR).

#### 5.2 Data Summary

We would be exploring a pre-loaded event log object (dataset) in bupaR called patients. patients is a fictitious event log about a hospital's processes involving patients’ admission. The dataset contains 5442 observations of 500 patients and 7 variables.

The table below shows a patients event log. Each row is an event that belongs to a **case** (*a* *patient*). Different events together can form an **activity instance**, or execution (e.g. event 3-4 belong to Registration 10*).*Each event in such execution will have a different transactional **lifecycle** status as start or complete. Also, there can be different instances of a specific **activity** (e.g., there are two surgeries). Furthermore, each event has a **timestamp**, indicating when it happened, and a **resource (employee)**, indicating who performed it [8].

The dataset is clean with no missing, invalid, or duplicate records. There are a total of 7 activities, carried out for 500 cases (patients), in 7 unique ways (traces) by 7 resources.



*Fig 1.: patients event log table*

**Data Dictionary:**

*handling*: the activities involved in the patients’ admission in the hospital (factor)

*patient*: patient id, unique for each patient (character)

*employee*: individual responsible for the instance of the process/activity (factor)

*handling\_id*: uniquely identifies instances of activity that is treated as different from others (character)

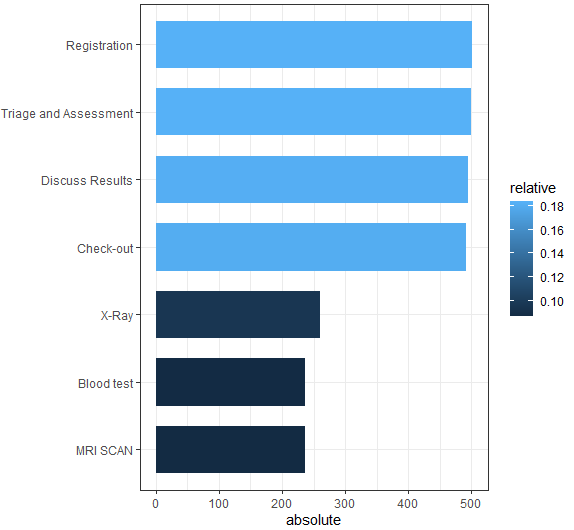
*registration\_type*: status as start or complete of the activity (factor)

*time*: timestamp at which logging was done (date time)

*.order*: a unique identifier for the whole dataset (integer)

#### 5.3 Explanatory Data Analysis

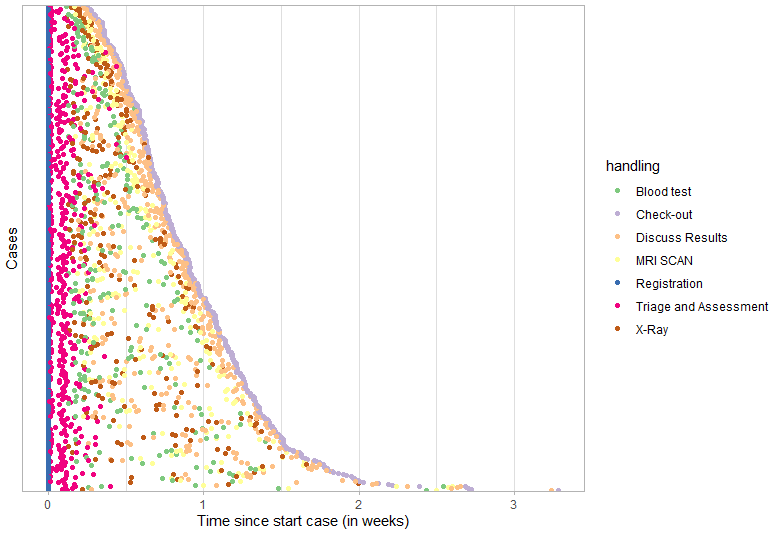
First, we will plot to understand what different activities are present and their frequency.



*Fig 2.: frequency of cases in each activity*

The plot shows that the highest frequency activity is Registration and Triage and Assessment, which is to be expected. Also, Discuss Results and Check-out has high frequency. X-Ray, Blood test, and MRI Scan have least frequency since not all patients go through these phases.

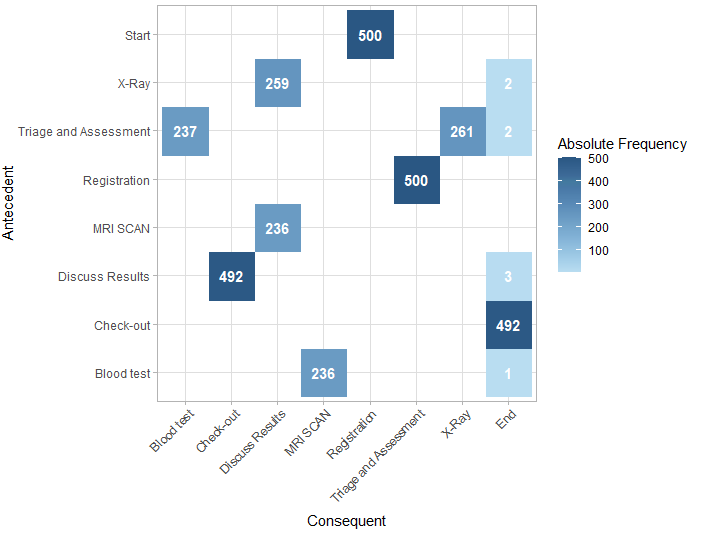
The dotted plot shows all the events by the relative time in weeks (where first event is counted as zero) and case identifier. The plot is sorted according to their overall cycle time.



*Fig 3.: Dotted chart of relative time in weeks and cases*

We see that there is a considerable variation in terms of cycle time. Furthermore, the chart suggests that there might be three distinct classes of cases: those that take no more than 1 week, those taking between 1 to 2 weeks, and a small class of cases taking longer than 2 weeks. Such an explorative inspection can provide a good basis for a more detailed analysis of the factors influencing the cycle time.

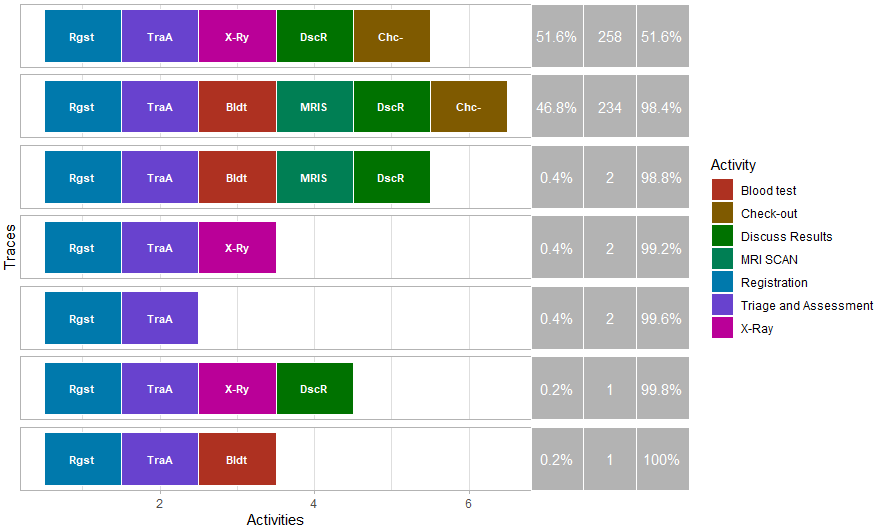
We visualize the precedence matrix which shows what steps tend to happen together.



*Fig 4.: Precedence matrix of each activity*

We see that Start point is always Registration whereas the End point is usually Check-out but not always. Registration and Triage and Assessment goes together in 500 cases (which is the total number of cases). Triage and Assessment goes together with Blood test in 237 times of cases, while with X-Ray in 261 times of cases. Discuss Results and Check-out goes together in 492 times of cases.

We now plot to get the frequent traces, i.e., activity sequences in the event log.



*Fig 5.: Frequent traces*

We see that the first two activity sequences cover 98.4% of the event log, meaning maximum of the cases follow these 2 paths of activities:

– Registration, Triage and Assessment, X-Ray, Discuss Results, Check-out

– Registration, Triage and Assessment, Blood test, MRI Scan, Discuss Results, Check-out

All the remaining sequence together covers only 1.6% of the cases which means a very few proportions of the cases follow the other paths in the hospital admission. These cases could be an open case or a closed without completion case since the traces do not contain Check-out.

###### 5.3.1 Summary statistics of processing time of each activity

We get the below summary statistics of each activity duration in minutes.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Activity | min | q1 | mean | median | q3 | max | std\_dev | iqr | relative\_freq |
| Registration | 49.3 | 124.2 | 165.2 | 162.8 | 204.1 | 338.0 | 57.2 | 79.8 | 0.184 |
| Triage and Assessment | 352.1 | 681.1 | 786.3 | 800.4 | 901.8 | 1128.1 | 165.5 | 220.7 | 0.184 |
| Discuss Results | 80.0 | 138.9 | 166.5 | 166.3 | 193.2 | 272.1 | 37.6 | 54.3 | 0.182 |
| Check-out | 40.0 | 96.7 | 123.7 | 124.3 | 148.3 | 233.7 | 37.2 | 51.6 | 0.181 |
| X-Ray | 137.7 | 233.2 | 290.7 | 287.5 | 338.9 | 490.1 | 76.8 | 105.7 | 0.096 |
| Blood test | 185.7 | 285.4 | 332.1 | 328.1 | 376.0 | 488.2 | 63.5 | 90.6 | 0.087 |
| MRI Scan | 149.4 | 216.4 | 248.9 | 245.4 | 281.8 | 355.4 | 44.1 | 65.3 | 0.086 |

*Table 1.: Summary Statistics of processing time of each activity*

We see that the Triage and Assessment usually takes much longer time than all the other activities. Whereas Registration, Check-out, and Discuss Results takes the least amount of time.

###### 5.3.2 Summary statistics of processing time of case

We get the below summary statistics for case duration in number of days.

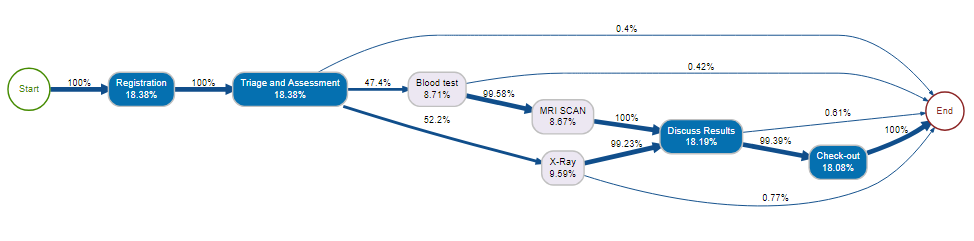
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| min | q1 | median | mean | q3 | max | std\_dev | iqr |
| 0.45 | 1.04 | 1.155 | 1.156 | 1.28 | 1.59 | 0.17 | 0.24 |

*Table 2.: Summary Statistics of processing time of case*

We see, on average, a case lasts for 1.156 days or 27.7 hours with a minimum of 0.45 days or 10.8 hours and a maximum of 1.59 days or 38.16 hours being invested.

#### 5.4 Process Model

Precedence relations between activities can also be shown with a process map. A process model combines all the cases in an event log to represent the mainstream behavior of the event log as a single process [9]. We create a process model or sometimes called, Petri net for the patients event log using the process\_map() function of processmapR library. We use the relative frequency in percentage to show the percentage of cases moving in that direction. Here 100% means the total number of cases, i.e., 500.



*Fig 6.: Process Model*

Each rectangle box denotes an activity and the color of boxes, blue and grey, denotes the frequency of the activity, which means blue colored activities is performed more often than the gray colored ones. Furthermore, the percentages inside the boxes shows how much is the contribution of that activity to happen out of all the 7 activities. Like Registration happened 18.38% times whereas MRI Scan happened only 8.67% of times in the process involved in patients event log.

The process starts from the Start where 100% of the cases moves to the Registration and then to Triage and Assessment, denoting that all the patients must have to go through these 2 steps during healthcare admission. The arrows represent the transition or next activity in the process. The activity of Triage and Assessment leads to a series of steps, whether to X-Ray, Blood test, as well as to the End with 52.2%, 47.4%, 0.4% of chances respectively, as each patient follows a different route, based on their condition and encounter. Also, it appears in some cases, that after the Blood test or X-Ray results, patients were then released without Check-out. Furthermore, in 99.58% of the patient cases getting Blood test, the patients go through MRI Scan.

As expected, there are high percentage of patient cases (~100%) where different tests (Blood test, MRI Scan and X-Ray) were utilized, as well as almost all patients proceed to Discuss results and Check-out but some cases where patients were never checked out. The process ends at the End point of the process model where all arrows converge. The bold lines show high values of relative frequency in percentage.

As can be seen from the model, some of the cases are not completed/recorded as it does not reach to check-out, such as Triage and Assessment, Blood test, Discuss results, or X-Ray directly reaching to the end of the process of admission without encountering the Check-out phase at 0.4%, 0.42%, 0.61% and 0.77% times, respectively. Therefore, there is a discrepancy in this process and thus needs to be worked upon. Also, we can compare the actual process and the logged process in the hospital admission to compare if there is indeed a discrepancy in the process or is it that some of the events are not logged properly in the database.

### **6 Conclusions**

In conclusion, process mining methods and algorithms appear effective in analyzing the available patients event log. Not only does the analysis of the event logs provide process mapping and process analysis, but also highlights areas in the clinical operations that may need further investigation. Modeling the patient pathway 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 [10]. In addition, it also appears that some data mining methods, such as creating a histogram, are 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 a 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 the dataset used is artificial, although real-life dataset like sepsis event log data also exists in the library. The other would be the use of only one dataset, which is one patients event log. Also, the library used for this study was bupaR while there are other libraries also available for process analysis and visualizations like subprocess and Rgraphviz. 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, to baseline and improve them. With limited software options available when analyzing event logs, it is difficult to accurately confirm the results, repeatedly.

Therefore, we recommend using the process mining techniques in the businesses as it helps in a 360-degree study of the ongoing processes and provides tools to study deeply each of the activities involved. The visualizations provided by such process mining techniques is quite useful and user-friendly. The businesses can make use of mining systems to get the best use of the information systems residing in the organizations and improvise upon the existing processes to serve their customers in the best possible way and maximize their profits.

References

[1] Alessandro Berti, Sebastiaan J. van Zelst, Wil M.P. van der Aalst - RWTH Aachen University, "Process Mining for Python (PM4Py): Bridging the Gap Between Process- and Data Science", 15 May 2019

[2] W. van der Aalst, Process Mining Data Science in Action. Springer-Verlag Berlin Heidelberg, 2016.

[3] Aguirre, S., Parra, C. and Sepúlveda, M. (2017) ‘Methodological proposal for process mining projects’, Int. J. Business Process Integration and Management, Vol. 8, No. 2, pp.102–113.

[4] Tugba Gurgen Erdogan and Ayca Tarhan, "A Goal-Driven Evaluation Method Based On Process Mining for Healthcare Processes" Received: 7 April 2018; Accepted: 27 May 2018; Published: 30 May 2018

[5] Gert Janssenswillen, Benoˆıt Depaire, UHasselt - Hasselt University, Faculty of Business Economics, Agoralaan, 3590 Diepenbeek, Belgium, Research Foundation Flanders (FWO), Egmontstraat 5, 1060 Brussels, Belgium, "bupaR: Business Process Analysis in R"

[6] Janssenswillen, G., Depaire, B., Swennen, M., Jans, M., & Vanhoof, K. (2019). bupaR: Enabling reproducible business process analysis. Knowledge-Based Systems, 163, 927-930.

[7] Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J.: shiny: Web application

framework for r, 2015. URL http://CRAN. R-project. org/package= shiny. R package version 0.11 (2015)

[8] [Gert Janssenswillen](https://www.r-bloggers.com/author/gert-janssenswillen/), September 29, 2017 , “bupaR: Business Process Analysis with R”

[9] Shannon Feeley, Ian Jacoway, Gina Rios, WORCESTER POLYTECHNIC INSTITUTE, April 27, 2017, “Process Mining the Credit Suisse Advisory Process”

[10] [Renee M. Hendricks](https://www.ncbi.nlm.nih.gov/pubmed/?term=Hendricks%20RM%5BAuthor%5D&cauthor=true&cauthor_uid=31632608), Published online 2019 Sep 19., “Process Mining of Incoming Patients with Sepsis”