Identification of differences in sub-population healthcare processes

Philipp Classen
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
p.r.classen@student.utwente.nl

ABSTRACT

Understanding care-path processes is important for optimising processes in hospitals. Process mining, which allows the mining of process models of care-paths, can help improve the care-paths followed by patients. Hospitals follow a general predefined procedure for treatment where deviations can occur which impact the health of a patient. Hence, there is a need that these processes are executed in the most efficient way. Process mining helps to identify which processes are being executed as well as bottlenecks and deviations. Analysing the deviations in a process and the frequency of activities allows optimisation and improvement in efficiency of a process. This research will use the MIMIC-III dataset and applies the L* Life-Cycle and PM^2 methodology onto it. This paper highlights differences in the care-path processes of the sub-populations 'insurance type' and 'age' with the same medical condition and shows that no differences can be found across the different genders.

Keywords

Process mining, healthcare, event log, sub-population, care-path, Medical Information Mart for Intensive Care III, MIMIC-III, ProM

1. INTRODUCTION

Hospitals around the world generate a tremendous amount of data for each and every patient. They perform an array of different processes that impact people's health. With an increasing demand for health care, medical centers are trying to optimise their processes. [12]

In a hospital, care-path defines the medical and management activities and decisions, and organization thereof, intended to carry out a care process in patient healthcare. These care-paths are generally described using a process model or workflow, which contain all possible paths of every patient. Typically, hospitals have to adhere to guidelines and predefined processes for the care-path of a patient. Despite clinical guidelines, deviations will occur as medical examiners have different levels of experience and patients might have multiple medical conditions. [6]

31^{th} Twente Student Conference on IT Jun. 5^{th}, 2019, Enschede, The Netherlands.

Consequently, there is a need for these processes to be understood and executed in the most effective and efficient way for all patients.

Process mining has come up as one of the disciplines that can discover, monitor and improve processes by utilizing event logs.[2] As hospitals follow processes that are highly complex and deviate between patients, more traditional techniques like drawing models or analysing by hand become unfeasible. Process mining allows to semi-automate process design and performance measurements that would not be achievable with related techniques like data mining.

This paper will focus on identifying differences and commonalities in care-path processes of sub-populations with the same medical condition. Moreover, optimisation approaches to improve process efficiency will be discussed.

First, this paper will provide background information on process mining and information about the MIMIC-III [5] dataset that is used. Afterwards, related work about process mining in health care will be discussed. Finally, the proposed methods for process mining and visualisation will be presented and results of the research will follow. At the end of the paper the results will be discussed and future research proposed.

2. RESEARCH QUESTION

Based on frequently posed questions by Mans et al. for process mining in healthcare [8] the following questions were identified as relevant:

RQ1: Which medical conditions provide data to identify sub-populations and what are their characteristics?

RQ2: In which steps of the care-path process can differences be found and what are those differences?

RQ3: Can selected sub-population processes be made more efficient?

Table 1. List of sub-populations

Sup-population	Description
Gender	Exploring the process difference between male and female patients
Age	Exploring the process difference between different age groups of patients
Insurance type	Exploring the process difference between different insurance types

3. BACKGROUND

3.1 MIMIC-III

MIMIC-III (Medical Information Mart for Intensive Care III) is a large, freely-available database comprising information relating to patients admitted to critical care units of the Beth Israel Deaconess Medical Center. The database includes vital signs, medications, laboratory measurements, observations and caregiver notes, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more. [5] The dataset version used in this research is MIMIC-III v1.4, released on 2 September 2016. [1]

3.1.1 Data format

MIMIC-III consists of 26 relational tables, where 16 tables contain timestamped event information, which are necessary to perform process mining. These tables are linked with an identifier. SUBJECT_ID refers to a unique patient, HADM_ID refers to a unique hospital admission.

3.2 Process Mining

Process mining is a discipline that allows the extraction of information from event logs. The goal of process mining is to identify, monitor, and improve real processes by extracting knowledge from the event logs. Existing techniques focus on three main areas: process discovery, conformance checking and enhancement. [1]

Process discovery addresses the identification of an process model from an event log, here the medical care-path. Noisy or missing data as well as complex processes, as they are found in hospitals, can make the process discovery a challenge.

Conformance checking takes as input a given process model and an event log. The goal of conformance checking is to verify if the modeled behavior represents the observed behavior in the event log.

Process enhancement aims to extend and improve an existing process model using information about the actual process in an event log. This can be applied to time events and can highlight bottlenecks in the process chain.

4. RELATED WORK

4.1 Process Mining

4.1.1 Process mining in health care

A case study in 2008 [11] first explored the applicability of process mining in healthcare to gain insight in complex hospital processes. However, the concern that process mining techniques have problems with unstructured processes, as they can be found in hospitals, was also raised in this study.

In a systematic review [3] of process mining in healthcare, covering 55 articles, the researchers highlighted the different focus of the research: 69% of the articles focused on process discovery and only 6% on process conformance. While the difference between treatment and organisational process research is small, only 17% focus on the clinical pathway.

Recent research of process mining in oncology using the MIMIC-III dataset [9] explored the applicability of process mining in healthcare specifically for cancer patients, and showed that complex hospital processes can be extracted, compared and bottlenecks can be detected.

[1]https://mimic.physionet.org/about/releasenotes/

4.1.2 Visualisation

Dixit [4] introduced the novel tool InterPretA, a plugin for ProM, for interactive process-oriented analysis. It supports exploratory analysis with different helicopter views on the process.

4.2 Impact

This research will apply the same general approach in line with most of the current research regarding process mining in healthcare, process discovery, conformance and enhancement, but will focus on the identification of differences in care-path processes of sub-populations. A start in researching care-paths has been made for patients with similar medical conditions but did not focus on differences in sub-populations on the same medical condition.[9]

Furthermore, for the identified differences, an attempt in optimising the care-path will be made to increase the efficiency of the care-path processes and thereby provide better healthcare for the patient.

5. METHODOLOGY & APPROACH

5.1 L* Life-Cycle Model

For this research we used the L* Life-cycle model, proposed by van der Aalst et al [13], for process mining of real-life processes. The L* Life-cycle model consists of five stages.

5.1.1 Stage 0: Plan and Justify

Stage 0 involved identifying research questions as a first step to guide the research forward. The research Questions in this paper were based on frequently posed questions by Mans et al. for process mining in healthcare.

5.1.2 Stage 1: Extract

Stage 1 involved the extraction of data from the dataset so that it can be combined to an event log that can later be used for process mining. Here the selection was made on a subset of care-path cases for patients with 'Diseases Of The Skin And Subcutaneous Tissue'.

To pre-process the data, the four data processing activities described by Maikel L. [14] for the PM^2 Methodology were used: creating views, aggregation events, enriching logs, and filtering logs.

The initial investigation of the MIMIC-III dataset revealed large numbers of cases and activities. To try to handle these large case numbers, different and more simplified levels of abstractions were created.

To be able to mine the dataset and focus on the differences of sub-populations, a general disease type was selected and only patients that got admitted into the hospital with at least one of the specified ICD9 codes were selected. In this research, patients with 'Diseases Of The Skin And Subcutaneous Tissue', identified with ICD9 code 680-709, were chosen. This was done firstly to limit the type of diseases for patients, as this research investigates differences of sub-populations of the patients and not differences of diseases. Secondly, we want to reduce the amount of events and traces of the dataset, as the number of events of cases ranged from 4 to roughly 600,000.

Furthermore, to create a more homogeneous sets, the data was split into different types of skin disease.

5.1.3 Stage 2 : Create control-flow model and connect event log

The extracted data from the MIMIC-III dataset were analysed using process mining techniques such as process discovery and conformance checking.

Process discovery was performed using the Inductive Visual Miner algorithm [7] . The Inductive Visual Miner was chosen due to being robust, able to handle noise well and focus on the main process instead of all the details.

5.1.4 Stage 3 : Create integrated process model

Stage 3 involves the enhancement of the process models discovered in the previous stage with additional perspectives. In this research, time was added as a perspective to investigate the activities with the longest waiting times in a process.

5.1.5 Stage 4: Operational support

As stage 4 of the L* Life-cycle model only works with concrete operational support such as detect, predict and recommend, it is outside of the scope of this research. Moreover, it is only possible to apply it to lasagna processes and not spaghetti processes. [13] Therefore is not possible to apply it to this research due to the unstructured processes.

5.1.6 Tools

The tools used in this research were PostgreSQL, with the PgAdmin4 interface, ProM 6.8 and 6.9 as well as Disco. PostgreSQL was used as database to store a local copy of the MIMIC-III database and to be able to run SQL-queries on the dataset for extraction and selection. ProM provided the process mining tools for discovery, conformance checking and enhancement of the process models. Disco was used for early and quick process discoveries to gain an understanding of the dataset.

6. RESULTS

6.1 Stage 0: Plan and Justify

The research questions for this research were formulated of frequently asked questions in process mining in healthcare. The research is of exploratory nature to investigate if differences in care-paths between sub-populations can be found.

6.2 Stage 1: Extract

In total 2872 patients were found to have at least one type of skin disease. All of those were selected for this research. Patients might have multiple types of skin disease and fall into multiple groups. The skin diseases were clustered into eight different groups based on their ICD codes and description.

The complete table containing all events for the patients had 44,204,892 rows. Using an event log of this size and complexity created a 'Spaghetti' model that can not be analysed in a useful manner and is not presented here.

To resolve this issue, a more simplified, filtered version of the event log was created based on the PM^2 methodology.

6.2.1 Filtering

Two types of timestamp were found in the dataset, all tables that only provided timestamps down to a day were excluded. These are the tables *callout, cptevents* and *prescription*. Entries without timestamp or activity were excluded from the dataset, as those two elements are necessary to perform process mining. Duplicates in the records were removed.

6.2.2 Enriching log

To be able to look at different levels of detail for a care-path, due to the size and complexity of the dataset, the event logs were extended with different levels of activity: Activity as the original activity as found in the dataset, category as the label of d_items and d_labitems, and the tablename as highest level of detail.

6.2.3 Creating views

Creating views was used to decide what level of detail was needed in the process discovery. The first level focused on the ins and outs of the hospital and provided insight into the high-level care-paths of a patient in the hospital.

To create more detailed models that go into the specific treatments of a patient, a combination of different levels of detail can provide an more detailed view, for example about the events happening during a ICU stay.

6.2.4 Aggregating events

For some traces in the dataset, multiple events with the same timestamp were found. This might happen due to data being imported into the hospital system from a subsystem and the import time was taken as the timestamp. Those events were mainly found in the more detailed views and were excluded from the subsets.

Table 2. Example Event Log

subject_id	hadm_id	activity	charttime
1	123456	ED registration	20/10/2101 17:09
1	123456	admission	20/10/2101 19:08
1	123456	ICU in	20/10/2101 19:10
1	123456	ED out	20/10/2101 19:24
1	123456	ICU out	26/10/2101 20:43
1	123456	discharge	29/10/2101 08:45

6.3 Stage 2: Create control flow

The second stage was to create a control flow model and connect it to the event log. For this, the ProM plugin Inductive Visual Miner was used to generate Petri Nets. The model was adapted and outliers and traces with irregular activities were filtered out to generalize and provide a better model.

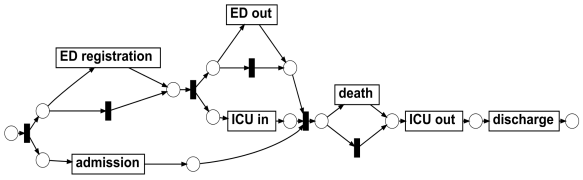

Figure 1. Petri net of full admission log

The Petri net in Figure 1 was created using the Inductive (visual) Miner [7] showing the model of 3153 admissions of 1585 patients that got admitted with ICD_CODE_681_2. This event log was also used to show the most frequent traces that occurred in the process.

The first similarities found across the sub-populations are that certain events will always happen, independent of sub-population assignment. The events *admission, ICU in, ICU out* and *discharge* are always experienced by each patient in the dataset.

This is not surprising, as every patient that enters the hospital ICU (Intensive Care Unit) has to be admitted into the hospital and ICU and also discharged from both as well. The events *ED out, ED in* and *death* are not

Table 3. Subsets of "Diseases Of The Skin And Subcutaneous Tissue"

#	Description	ICD9 code	Patient #	Admission #
1	Caruncle and Furuncle	680	10	21
2	Cellulitis and abscess	681-682	1585	3153
3	Lymphadentis, Impetigo, Pilonidal cyst and other local infections	683-686	88	144
4	Dermatosis and dermatitis	690-694	826	1341
5	Erythematous, psoriasis, lichen and pruritus	695-698	551	969
6	Corns, callosities, hypertrophic and atrophic conditions of skin	700-702	16	64
7	Diseases of Nail, hair and follicles, sebaceous glands and sweat glands.	703-706	34	101
8	Chronic ulcer of skin, urticaria, and other disorders of skin and subcutaneous tissues.	707-709	627	1565
All	(680-709)		2872	5134

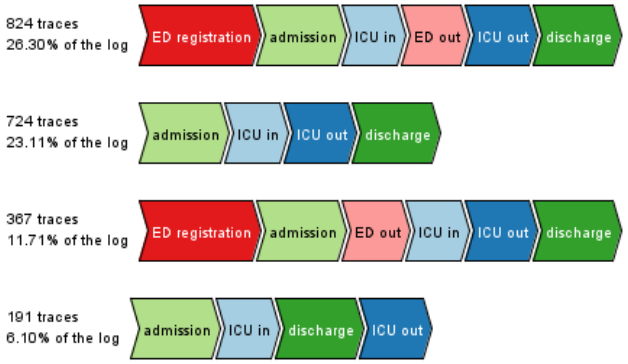

Figure 2. The four most followed trace covering 67,22%

experienced by all patients. *ED in* and *ED out* are only followed by patients that enter the hospital through the Emergency Department(ED) and *death* only happened to 175 patients admitted with ICD9 681-682, and is a very rare occurrence compared to other events.

The sub-population of insurance type *Self Pay* was not used in this research as the number of patients admitted into the hospital was too low in comparison to other sub-populations in the insurance group.

For the age sub-population, the groups *neonate* and *child (1-20)* were excluded from the research as well. Neonates were excluded due to the assumption that there are already big differences in treatment for a neonate compared to a normal adult, which would influence the result. The group of 1-20 year old was excluded due to the low amount of patients in that group. See Table 4 and Table 5.

Table 4. Insurance type

Insurance type	Patients #
Government	107
Self Pay	25
Private	1033
Medicare	1539
Medicaid	326
Total	3030

To identify which sub-population would provide the best model for all other sub-populations, an event log and a Petri net were created. With the event log and the Petri net, a model-log matrix was created that shows the fitness and precision of each log replayed on each model. See Table 6, Table 7 and Table 8.

Table 5. Age

Age	Patients
Neonate (< 1)	181
Child/teen (1-20)	8
Young adult (20-40)	254
Adult (40-64)	1192
Elder (65-89)	1128
Old Elder (>89)	109
Total	2872

Fitness: A model with good fitness represents the behavior seen in the event log. A model has a perfect fitness (1.0) if all traces in the log can be replayed by the model from beginning to end.

Precision: A model with a good precision does not allow behavior completely unrelated to what is found in the log. A very low precision would indicate a under-fitted model and a perfect precision(1.0) would be an over-fitted model that only allows what was seen in the model.

In Tables 6, 7 and 8 we can see that all logs and models have at least a fitness of 0.862 or higher. This indicates that the models generally allow all the events to happen that were seen in the event logs and shows that the events between the sub-populations are similar.

For precision we find bigger differences between the different groups. The *male* and *Medicare* models show the biggest differences compared to the other models, with precision values as low as 0.558 for the government log replayed on the male model. These low values show that these models allow behavior relatively unrelated to what was seen in the event log and do not make a great generalized model for the sub-populations.

The models generated from the sub-population of *Age>89* and *Medicaid* show the highest precision overall, for some logs even a value of 1 as well as high values of fitness as well. This makes them good candidates to use as general models to represent the process. The models allow the behavior seen in the log and not more while also being able to replay almost all traces from it.

6.4 Stage 3: Create integrated process model

To investigate whether there are time differences and commonalities in the paths of different sub-populations, process mining log replay was used to highlight the average waiting time between different events. For this, the event logs of the sub-population groups were put in the "Multi-Perspective Process Explorer" plugin [10] and were replayed on the model generated by the sub-population *Medicaid*.

In the following graphics the first number on the edges show the percentage of traces that have taken this path and the numbers in brackets show the average waiting time between two events.

Figure 3 shows the performance view, from the "Multi-Perspective Process Explorer", of the full log replayed on the Petri net.

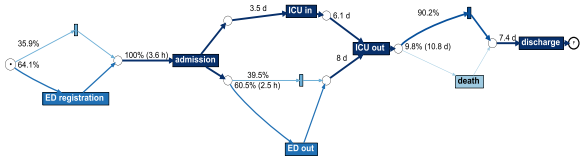

Figure 3. Full log replayed on Petri net.

6.4.1 *Gender sub-population*

This was performed for all the different sub-population groups, to highlight the paths with the longest waiting times. Figure 4 & Figure 5 show the performance view of the *male* and *female* sub-populations. No noticeable differences between the sub-populations of *female* and *male* patients are seen here. The biggest differences, of 0.6 days, that can be noticed between *male* and *female*, is visible at the death event. This difference is most likely due to the few *death* events in the subset overall, 85 patients of the male and 90 patients of the female sub-population, and not due to differences of the sub-populations.

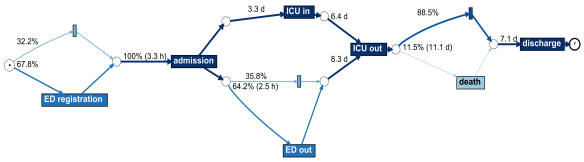

Figure 4. Female Log replayed on Petri net.

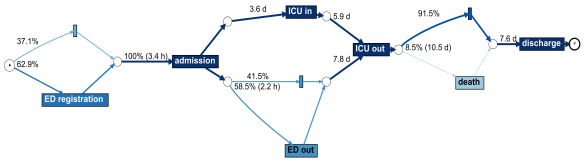

Figure 5. Male Log replayed on Petri net.

6.4.2 *Insurance sub-population*

When investigating the differences in waiting time for different types of insurance, more differences are shown. The performance view showed that patients with a *Medicare* insurance spend on average only 4.9 days in the Intensive Care Unit and leave the hospital 6.7 days later already. Patients with other types of insurance spend around 7-8 days in the Intensive Care Unit and leave the hospital 8-10 days later. See Figure 6. This shows that a *medicare* insured patient spend on average 11.6 days in the hospital from entering the ICU to leaving the hospital, compared to patients with different insurance types, who spend 16.5 days in a hospital form entering the ICU to leaving. This means they spend on average 40% more time in the hospital.

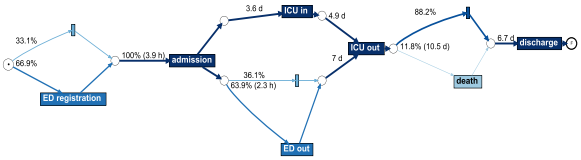

Figure 6. Medicare log replayed on Petri net.

6.4.3 *Age sub-population*

For the different age groups only one difference stands out. Patients of the age group 89+ spend less time in the intensive care unit compared to all other age groups. Patients from the group of 89+ years spend on average 3.6 days in the ICU and leave the hospital 4.7 days later. All the other age groups spend on average 5 days in ICU. Patients between 20 and 40 years leave the hospital on average 10 days after being discharged from the ICU. See Figure 7 and Figure 8. This shows that a 89+ patient spend on average 8.3 days in the hospital from entering the ICU to leaving the hospital, compared to a young adult (20-40), who spends 15 days in a hospital form entering the ICU to leaving. This makes a time difference of 6.7 days and is almost an increase of 100%.

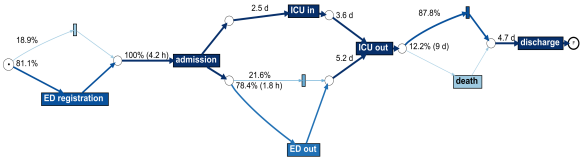

Figure 7. Age 89+ log replayed on Petri net.

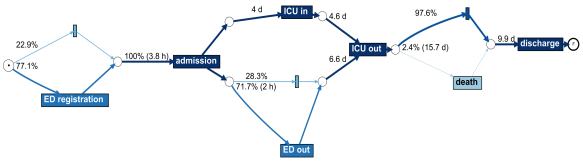

Figure 8. Age 20-40 Log replayed on Petri net.

To investigate the differences found, a more detailed model needs to be created based on the sub-populations. An attempt to mine such a process model was made during the research but dismissed due to computational limitations of the device used.

7. DISCUSSION

The L* Life-Cycle Methodology by van der Aalst provides a helpful methodology for the structuring and preprocessing of real world data.

7.1 Sub-populations

For this research, only patients with a 'Disease of the Skin or Subcutaneous Tissue' were used. This selection was based on available data and not already performed research. Further, a too big subset caused trouble with the process mining tools and a too small subset would not require process mining so differences might not be easily found in the data.

This means the model generated in this research will only hold for this specific type of disease but the same procedure can be performed on other diseases for further investigation.

To limit the scope of the research, only a small part of the different sub-populations were selected. We only tested differences for gender, insurance type and age but more sub-populations can be found in the dataset. Sub-population of, for example ethnicity or religion can be found in the dataset but for those two sub-populations not all patient provided data as a patient can abstain from answering. For the sub-populations gender, age and insurance type the patient has to provide the data to the hospital.

7.2 Differences

While we have not found any differences between male and female patients on the here selected disease, there still might be differences between the genders found somewhere else. At a first glance this is a good result as there should not be any differences in care-path for male and female patients just based on their gender. But other diseases could highlight differences between male and female. These differences in care-path could be based on different treatment paths performed for male and female based on the physical differences that require different treatments.

In this research we have found that patients with a *Medicare* insurance spend less time in the hospital than other insurance types. Medicare is a special type of insurance that only people who are older than 65, have certain disabilities or have ESRD or ALS are eligible for. No other type of insurance left the hospital that fast.

There might be multiple explanations based on the insurance or age. Medicare is a US federal governed health insurance that most people above 65 are automatically enrolled in. This insurance is covered by payroll taxes, including medicare and social security taxes, and basic coverage cost nothing for those who payed into the system for 10+ years. This includes coverage of hospital stays. Hence, the shorter the stay of patient in the hospital, the lower the amount the federal insurance has to cover for those patients.

While the differences found are only for a specific disease, the same measurements should be performed for other diseases to investigate if the same differences can be found for other diseases and if a generalization could be made about *medicare* patients and the time they spend in the hospital.

As discussed above about eligibility for a type of insurance, that is based on age, we found that patients above the age of 89 leave the hospital twice as fast as young patients. This can have multiple reasons. Those two differences might also be correlated. Patients eligible for medicare insurance are generally over 65 and there could be a big overlap between patients over 89 and patients with a medicare insurance. This could explain why both groups show a similar behavior and leave the hospital faster than their comparative groups.

7.3 Optimisation

While this research only investigated whether there are differences between selected sub-populations, it was not possible to provide optimisations for the care-path processes for the differences found. Due to the nature of the data, it is not possible to connect a treatment event to a specific disease. As patients are admitted into the hospital with multiple ICD9 codes, it is not clear for which disease a certain treatment or measurement was performed. This makes it difficult to improve the treatment for a specific injury or sub-population without medical knowledge of the disease and their treatment process.

7.4 Process mining vs. Data mining

While it was mentioned at the beginning of the paper that data mining is not feasible for these complex processes, throughout the research it became apparent that process mining also has its limitations. Process mining can be used to investigate differences in care-paths but reaches its limits for complex spaghetti processes, as they are found in a hospital. Using data mining techniques to prepare the dataset before applying process mining could help with reducing complexity of the dataset and make process mining easier.

7.5 Issues with ProM

In an attempt to generate a more detailed process model, the hardware limitations of the researchers device as well as the limitations of the process mining tool ProM were reached. While ProM theoretically is not limited in scale and model complexity, subsets from the MIMIC-III database are still too big for ProM on a normal device. The physical memory as well as computational speed of the hardware to mine the process models are the limiting factors here. Moving the process mining onto an external machine with better optimisation for data and process mining could improve the performance.

Furthermore, other mining techniques, like data mining, might be better to understand the data and find specific treatment process cluster, that could be investigated for differences.

7.6 Future Work

Future research should investigate whether more detailed models can be generated for these sub-populations with the process mining tool ProM. It should also be combined with the knowledge of a medical expert in the field to gain a general understanding of the treatment process for these diseases to create a model beforehand.

More research should also be done with other mining techniques like data mining and if those techniques can provide insight into the process as well.

8. CONCLUSION

In this research we investigated if differences of care-paths can be found between sub-populations in a hospital. We conducted process discovery to identify process models for the different sub-populations and performed fitness and precision measurements. Afterwards event log replay was performed to show whether time differences are present. First we explored the MIMIC-III dataset and followed the PM^2 methodology to prepare our data for process mining. This was followed by process discovery and conformance checking via ProM.

As a result of this research, we conclude that there are no noticeable differences between the male and female population. However, differences were found in the sub-populations of *age 89+* and insurance type *medicare* for patients with 'Diseases Of The Skin And Subcutaneous Tissue'.

9. REFERENCES

[1] W. M. P. Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, volume 136. Jan. 2011.

[2] W. v. d. Aalst. *Process Mining: Data Science in Action*. Springer-Verlag, Berlin Heidelberg, 2 edition, 2016.

[3] E. Batista and A. Solanas. Process mining in healthcare: A systematic review. 2019.

[4] P. Dixit, H. Garcia Caballero, A. Corvo, B. Hompes, J. C. A. M. Buijs, and W. M. P. Aalst. Enabling Interactive Process Analysis with Process Mining and Visual Analytics. pages 573–584, Feb. 2017.

[5] A. Edward William Johnson, T. Joseph Pollard, L. Shen, L.-w. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony G. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, May 2016.

[6] P. Homayounfar. Process mining challenges in hospital information systems. page 6, 2012.

[7] S. J J Leemans, D. Fahland, and W. M. P. Aalst. Process and Deviation Exploration with Inductive Visual Miner. volume 1295, page 46, Jan. 2014.

[8] U. Kaymak, R. Mans, T. v. d. Steeg, and M. Dierks. On process mining in health care. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1859–1864, Oct. 2012.

[9] A. Kurniati, G. Hall, D. Hogg, and O. Johnson. Process mining in oncology using the MIMIC-III dataset. volume 971, 2018.

[10] F. Mannhardt, M. de Leoni, and H. A. Reijers. The Multi-perspective Process Explorer. volume 1418, Aug. 2015.

[11] R. Mans, M. Schonenberg, M. Song, W. Van Der Aalst, and P. Bakker. Process mining in healthcare - A case study. volume 1, pages 118–125, 2008.

[12] R. S. Mans, W. M. P. van der Aalst, R. J. B. Vanwersch, and A. J. Moleman. Process Mining in Healthcare: Data Challenges When Answering Frequently Posed Questions. In R. Lenz, S. Miksch, M. Peleg, M. Reichert, D. Riaño, and A. ten Teije, editors, *Process Support and Knowledge Representation in Health Care*, Lecture Notes in Computer Science, pages 140–153. Springer Berlin Heidelberg, 2013.

[13] W. van der Aalst. Process mining: discovering and improving Spaghetti and Lasagna processes. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 1–7, Paris, Apr. 2011. IEEE.

[14] M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst. PM ^{2} : A Process Mining Project Methodology. In J. Zdravkovic, M. Kirikova, and P. Johannesson, editors, *Advanced Information Systems Engineering*, volume 9097, pages 297–313. Springer International Publishing, Cham, 2015.

APPENDIX

A. TABLES

Table 6. Model Log Matrix Gender, F - Fitness, P - Precision

Gender models	Full Model		Male Model		Female Model	
	F	P	F	P	F	P
Full Log	0.916	0.887	0.919	0.760	0.927	0.953
Male Log	0.917	0.911	0.922	0.743	0.928	0.854
Female Log	0.911	0.816	0.917	0.684	0.929	0.926
Age 20 - 40	0.926	0.761	0.927	0.585	0.941	0.821
Age 40 - 64	0.918	0.908	0.923	0.731	0.927	0.926
Age 65 - 89	0.915	0.851	0.921	0.678	0.927	0.801
Age >89	0.913	0.791	0.920	0.608	0.918	0.836
Government	0.922	0.745	0.923	0.558	0.935	0.797
Medicare	0.912	0.915	0.919	0.758	0.924	0.878
Medicaid	0.929	0.781	0.931	0.614	0.945	0.832
Private	0.914	0.863	0.916	0.654	0.931	0.927

Table 7. Model Log Matrix Age group, F - Fitness, P - Precision

Age models	Age 20-40 Model		Age 40-64 Model		Age 65-89 Model		Age >89 Model	
	F	P	F	P	F	P	F	P
Full Log	0.912	0.874	0.906	0.885	0.912	0.896	0.863	1
Male Log	0.915	0.850	0.911	0.792	0.916	0.872	0.863	1
Female Log	0.911	0.831	0.904	0.837	0.911	0.855	0.868	1
Age 20 - 40	0.926	0.757	0.925	0.681	0.926	0.737	0.891	0.960
Age 40 - 64	0.917	0.852	0.913	0.860	0.935	0.922	0.866	1
Age 65 - 89	0.914	0.784	0.906	0.723	0.942	0.880	0.866	1
Age >89	0.911	0.784	0.906	0.733	0.931	0.849	0.893	0.954
Government	0.922	0.742	0.920	0.653	0.938	0.796	0.862	0.942
Medicare	0.911	0.874	0.904	0.813	0.936	0.946	0.868	1
Medicaid	0.928	0.781	0.926	0.732	0.948	0.871	0.910	0.903
Private	0.912	0.824	0.908	0.838	0.937	0.904	0.897	0.902

Table 8. Model Log Matrix Insurance type, F - Fitness, P - Precision

Insurance Models	Government Model		Medicare Model		Medicaid Model		Private Model	
	F	P	F	P	F	P	F	P
Full Log	0.916	0.830	0.946	0.779	0.911	1	0.927	0.913
Male Log	0.919	0.809	0.946	0.764	0.914	1	0.928	0.864
Female Log	0.916	0.793	0.947	0.703	0.911	0.999	0.929	0.885
Age 20 - 40	0.927	0.709	0.949	0.623	0.921	0.873	0.941	0.833
Age 40 - 64	0.920	0.810	0.944	0.738	0.913	0.999	0.927	0.887
Age 65 - 89	0.920	0.751	0.946	0.682	0.916	0.939	0.927	0.815
Age >89	0.919	0.723	0.937	0.642	0.914	0.856	0.918	0.847
Government	0.923	0.678	0.946	0.603	0.915	0.800	0.935	0.811
Medicare	0.916	0.830	0.944	0.779	0.911	1	0.924	0.887
Medicaid	0.930	0.731	0.957	0.658	0.922	0.869	0.945	0.843
Private	0.913	0.802	0.945	0.746	0.909	0.995	0.930	0.893