Process mining in oncology using the MIMIC-III dataset

IOP Conf. Series: Journal of Physics: Conf. Series **971** (2018) 012008 doi:10.1088/1742-6596/971/1/012008

Process mining in oncology using the MIMIC-III dataset

Angelina Prima Kurniati[1,2], Geoff Hall[3,4], David Hogg[1], Owen Johnson[1]

[1]School of Computing, University of Leeds, UK
[2]School of Computing, Telkom University, Bandung, Indonesia
[3]School of Medicine, University of Leeds, UK
[4]Leeds Institute of Cancer and Pathology, St James's University Hospital, Leeds, UK
Corresponding author's e-mail address: scapk@leeds.ac.uk

Abstract. — Process mining is a data analytics approach to discover and analyse process models based on the real activities captured in information systems. There is a growing body of literature on process mining in healthcare, including oncology, the study of cancer. In earlier work we found 37 peer-reviewed papers describing process mining research in oncology with a regular complaint being the limited availability and accessibility of datasets with suitable information for process mining. Publicly available datasets are one option and this paper describes the potential to use MIMIC-III, for process mining in oncology. MIMIC-III is a large open access dataset of de-identified patient records. There are 134 publications listed as using the MIMIC dataset, but none of them have used process mining. The MIMIC-III dataset has 16 event tables which are potentially useful for process mining and this paper demonstrates the opportunities to use MIMIC-III for process mining in oncology. Our research applied the L* lifecycle method to provide a worked example showing how process mining can be used to analyse cancer pathways. The results and data quality limitations are discussed along with opportunities for further work and reflection on the value of MIMIC-III for reproducible process mining research.

1. Introduction

Cancer can affect any part of the body [1] and is recognized as a large group of diseases with at least 65 recognised types of cancer [2]. The complex nature of this disease makes choices in cancer care pathways particularly challenging. Process mining offers the opportunity to develop a deeper understanding of this complexity and, if applied to cancer patient records, may help improve cancer care pathways and outcomes for cancer patients.

Process mining [3] is an emerging approach for discovering and analysing business process models which uses data extracted from the event logs in information systems. An event log is a record of timestamped activities automatically generated by the system as it is used to support the business activities of an organisation. These records can provide insights into the real-world effectiveness of business processes. Process mining has been applied to the analysis of healthcare processes with the aims of improving quality of care, patient safety, and optimization of resources [4]. Process mining is especially useful to analyse highly complex and flexible patient care processes (care pathways), as happen in cancer patient treatments.

Process mining has been used in healthcare to describe what happened, why it happened, what will happen, and what is the best that can happen [5]. Rojas et al. [6] reviewed previous studies using process mining in healthcare, and found that the most process mining case studies in healthcare were in oncology. A previous study by the authors [7] found 37 peer reviewed papers describing process mining based research in oncology with a regular complaint being the limited availability and accessibility of

suitable fine grained datasets with suitable information for process mining. Most of the papers (24 of 37) focused on gynaecological cancer, mainly because there was a Business Process Improvement (BPI) challenge held using this dataset [8]. We conclude that the availability of data for researchers is a key enabler for process mining research. Specifically, the availability and accessibility of sufficiently large volumes of detailed patient data is due to the sensitivity and confidentiality of such personal data.

Publicly available datasets are one option and this paper describes the potential to use MIMIC-III for process mining in oncology. The MIMIC-III dataset is an open access dataset from a hospital in the USA with a large number (n=46,520) of de-identified patient records. This paper presents a worked example to explain how MIMIC-III can be used to analyse patient data using process mining tools. We describe how selection criteria were used to construct event logs from cancer patient treatment records and then analysed to gain insight into treatment pathways. Our experiments are reproducible using the available information from the MIMIC-III database and all SQL queries, figures, resulting process models and supporting resources have been made available in a GitHub repository.

2. Background

2.1. Healthcare processes

The processes during a patient's time in a hospital consists of many different activities, including administrative (admission, discharge, transfer to a ward, etc.) and clinical activities (triage, test and scans, diagnosis, therapy, etc.) [9]. These activities are performed by different clinical roles (doctors, nurses, technical specialist, etc.) and vary from one healthcare organization to another [5]. Healthcare processes are recognized as nontrivial because the steps involved are nonlinear, complex and unpredictable, such processes do not always follow standard sequences [10].

Many healthcare organizations are now using electronic health record (EHR) systems to record administrative and clinical information about their patients and track the treatments provided. These EHR systems evolved from paper-based physician notes and the requirement to structure records more formally has increased as health organisations have grown in size and complexity [11]. Patient-level information, including demographic data and some clinical information (e.g. allergies, long-term conditions) is supplemented by time-stamped records recording observations, diagnosis, prescriptions, treatment and administrative processes such as admission and discharge [12]. Event data will generally be a mixture of coded variables and natural language text logged against the date, time, user id and type of event. These EHR systems therefore contain longitudinal data that can be explored using process mining techniques [4] and, as EHR systems mature, the opportunities to find and analyse process information about treatment pathways are growing.

Our paper focuses on cancer patient treatment and the sequence of administrative steps in the care pathways. Because we work only in the administrative steps, we hope that the approach and the data are appropriate for more general use.

2.2. Process mining

Process mining is an emerging research discipline which combines computational intelligence, data mining, process modelling and analysis approaches. The idea of process mining is to discover, monitor, and improve real processes by extracting knowledge from the event logs that can be extracted from business information systems [3]. Process mining is complementary with more traditional process modelling performed by business analysts.

There are three types of process mining: discovery, conformance checking, and enhancement [13]. Discovery takes an event log and creates a graphical business process model. Conformance checking can be used to check if reality, as recorded in the log, conforms to the model and vice versa. In process modelling a process can be represented as a directed graph structure with nodes representing transitions (i.e. events that may occur, e.g. chemotherapy, follow up) and places (i.e. conditions). The directed arcs determine which places are pre- and/or post conditions for which transitions. This structure for modelling a business process can be shown graphically following notations such as Petri nets [14] or

Business Process Modelling Notation (BPMN) [15]. The model can be used for further analysis, such as detecting deviations, deadlocks, and repairing or enhancing the model. Enhancement extends or improves the process model using additional information, such as resources, decision rules, and performance data [16]. In our literature review, 36 out of 37 studies analysed the control-flow (the ordering of activities) and performance perspectives. Those studies implemented process discovery and conformance checking to check model conformance against the event log. Most of the studies (24 of 37) used the ProM software tool framework (www.promtools.org). This paper presented the implementation of a methodology called the L* lifecycle model [16] covering process discovery and conformance checking in both control-flow and time perspectives.

2.3. The MIMIC-III Dataset

Medical Information Mart for Intensive Care III (MIMIC-III) is a database comprising EHR information related to patients admitted to critical care units at the Beth Israel Deaconess Medical Centre (BIDMC), in Boston, USA. The dataset version used in this research is MIMIC-III v1.4, released on September 20th, 2016 [17]. It contains data such as: vital signs, medications, laboratory measurements from within the hospital (i.e. in-patient) and from clinics (i.e. out-patient), charted observations during a patient's stay in the intensive care unit, and de-identified notes regarding the patient's stay, including nursing notes, physician notes and discharge summaries [18].

MIMIC-III consists of 26 relational tables, where 16 of them contain timestamped event information. Tables are linked by identifiers: SUBJECT_ID refers to a unique patient and HADM_ID refers to a unique admission. This study will focus on using the 16 event tables, but other tables provide supporting information. For example, when we used `chartevents` table, we would need to refer to `d_item` table to get the label of `item_id` specified in the `chartevents` table. Diseases and procedures in the MIMIC-III are encoded using the International Classification of Diseases version 9 (ICD-9) codes, and the mapping can be found in `diagnoses_icd` and `procedures_icd` tables.

Time in the MIMIC-III database is stored with one of two suffixes: TIME (down to the minute) and DATE (down to the day). Most data are recorded with a time indicating when the event took place (CHARTTIME) and when it was validated (STORETIME). In this study, the event logs were created using CHARTTIME attributes, as this is the best match to the time of actual measurement. All patient data in the MIMIC-III database has been de-identified and all dates have been randomly shifted to the future so that dates are internally consistent for the same patient but inconsistent across patients. The handling of dates and times presents issue for process mining which we will discuss.

3. Methodology

This study was exploratory in nature, with the goal of understanding cancer treatment processes in the MIMIC-III dataset as a publicly available data. The methodology in this research is the L* lifecycle model, suggested by van der Aalst et al [16] for typical process mining projects. Stage 1 (Extraction) was enriched with adoption of *data processing* stage in Process Mining Project Methodology (PM^2)[19]. Note that Stage 1 of the L* life-cycle model does not detail the extraction activities. The PM^2 does provide helpful guidance and these were added in Stage 1. Note also that the Stage 4 of the L* life-cycle model covers explicit operational support and was therefore beyond the scope of this particular study.

In Stage 0 (Plan and justify), planning involved identifying research questions as a starting point for investigation. We drew our research questions from our literature review of cancer pathways.

Stage 1 (Extract) involved extraction to build the log by applying the selection criteria to collect records of treatment cases for patients diagnosed with cancer. Our approach included four preprocessing activities described in the PM^2: *creating views, aggregating events, enriching logs, and filtering logs*. Initial investigations revealed a large number of cases and our approach was to create simplified extracts of the dataset at three different level of abstractions and also select specific cancer types to ensure more homogeneous subset of cases.

In Stage 2 (Create control-flow model and connect event log), the extracted data subsets were analysed using several process mining approaches. This was related to discovery and conformance

checking. In this paper, discovery was performed using the Inductive Visual Miner [20] algorithm. Inductive Visual Miner was used since it is robust, user friendly, and feature rich, can deal with noise and exceptions, and enabled the focus to be on the main process flow instead of on every detail of the behaviour appearing in the process log.

Stage 3 (Create integrated process model) extended the process models discovered in the Stage 2 with an additional perspective. The analysis was performed using Petri Net as a process modelling standard, to investigate the activities with the longest waiting times.

The tools used in this study were PostgreSQL (through PgAdminIII graphical interface), Python, and ProM 6.5.1. PostgreSQL was used as the database management system which allowed SQL-based queries though PgAdminIII to extract and select data from MIMIC-III database. Python was used to create more structured way of data processing. The ProM provided process mining toolset for discovery, conformance checking, and enhancement of process models. The SQL queries and resulting models are available for reuse in Github repository (https://github.com/angelinast3/mimic3cancerpromin).

4. Results

4.1. Stage 0: Plan and Justify

Planning for this study was started with a list of frequently posed questions suggested by Mans et all for process mining in healthcare [21]. Those questions were adapted in this study:

Q1. What are the most followed paths and what exceptional paths are followed?
Q2. Are there differences in care paths followed by different patient groups?
Q3. Where are the long waiting time activities in the process?

4.2. Stage 1: Extract

The extract stage in this study started with selecting records of patients diagnosed with cancer. The cancer codes in ICD-9 are 140x-239x [22], and can be grouped based on the cancer types. These codes can be found in `diagnoses_icd` table.

In total, 7,361 patients were found to have at least one diagnoses related to cancer. All of these were selected for this study. Note that a patient might have more than one type of cancer, which make this patient fall into more than one group. Total patients in this study were recorded against 13 cancer types, with the highest number in cancer type 7 (2,846 patients), cancer type 2 (1,400 patients), and cancer type 8 (1,110 patients). The results are summarized in Table 1.

Table 1. Summary of Cancer Types

#	Description	Patients	Admissions
1	Malignant neoplasm of lip, oral cavity, and pharynx (140-149)	87	135
2	Malignant neoplasm of digestive organs and peritoneum (150-159)	1400	**2012**
3	Malignant neoplasm of respiratory and intrathoracic organs (160-165)	1056	1540
4	Malignant neoplasm of bone, connective tissue, skin, and breast (170-175)	279	337
5	Kaposi's sarcoma (176-176)	14	19
6	Malignant neoplasm of genitourinary organs (179-189)	722	1025
7	Malignant neoplasm of other and unspecified sites (190-199)	2846	**3950**
8	Malignant neoplasm of lymphatic and hematopoietic tissue (200-209)	1110	1692
9	Neuroendocrine tumors (209-209)	26	38
10	Benign neoplasm (210-229)	1215	**2036**
11	Carcinoma in situ (230-234)	47	65
12	Neoplasms of uncertain behavior (235-238)	675	1065
13	Neoplasms of uncertain nature (239)	60	105
	Any type of cancer (140 - 239)	7361	10857

Extraction of all cancer patient data was done by selecting records from each of the 16 event tables in MIMIC-III. A summary of the extracted records is presented in Table 2.

Table 2. Summary of Extracted Tables

#	Table Name	Patients	Activities	Rows
1	admissions	7361	8	35843
2	callout	4197	6	27402
3	chartevents	7359	2580	38766594
4	cptevents	3327	1	19310
5	datetimeevents	5648	148	925542
6	icustays	7345	2	22976
7	inputevents_cv	3924	756	1833886
8	inputevents_mv	3850	251	664209

#	Table Name	Patients	Activities	Rows
9	labevents	7351	556	6912233
10	microbiologyevents	3457	47	11670
11	noteevents	5351	584	129712
12	outputevents	7278	415	824665
13	prescriptions	6900	2697	685648
14	procedureevents_mv	3853	114	53440
15	services	7357	18	15657
16	transfers	6902	8	29708
	Total	7361	7929	50958495

The resulting `allevents` table contained 50,958,495 rows. Using this table directly for process mining resulted in a "spaghetti" process model which is impossible to analysed and it is not presented here. Further transformation and creation of a simplified version of the dataset was essential.

Transformation was performed in several steps, as shown in Fig. 1. A transaction table was created consisting of `subject_id`, `activity`, `category`, `tablename`, `charttime` and records were inserted from all tables extracted before.

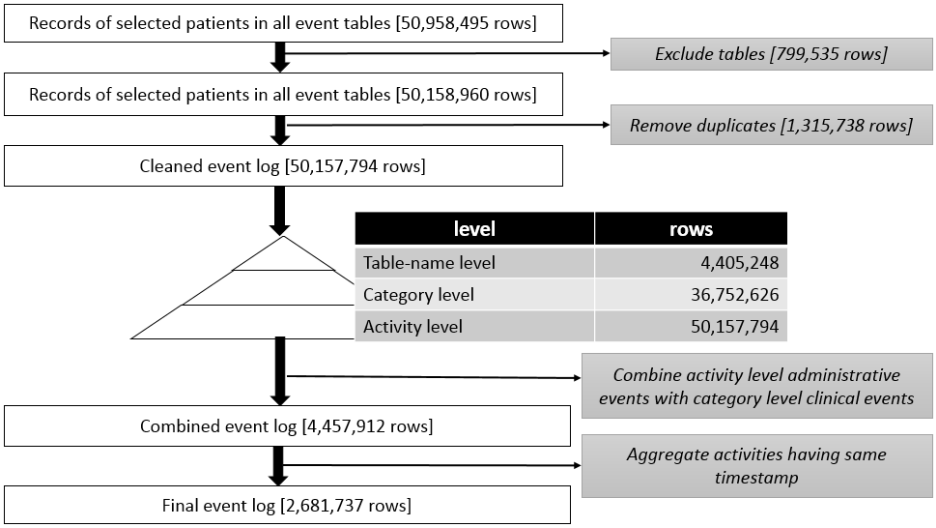

Fig. 1. The sequence of data extraction and transformation in Stage 1

Each row in the table was transformed into an event log record with the following preprocessing based on PM^2 methodology:

1) Filtering log

a. Three tables were excluded because they provide timestamps down to the day only. These were `callout`, `cptevents`, and `prescription` tables.
b. Records which did not have a timestamp or an activity name were excluded.
c. Duplicate records were reconciled by keeping one record and deleting other duplicate records. For example: `admission` in the `admissions` table, `Hospital Admit Date` in `datetimeevents` table, and `admit` in `transfers` table are refering to the same event.

2) *Enriching log* by creating three levels of details with three different values for activity labels, which are the original activity names (activity level), category names from `d_items` and `d_labitems` tables (category level), and table name (table-name level). Those three different levels are needed to make it possible to analyse the data in three different level of details, based on information available in the dataset.

3) *Creating views* was done by deciding which level of detail needed in the next stage. This was done based on two general types of events recorded in the tables, which are administrative and clinical events. In the MIMIC-III, administrative events are recorded in `admissions`, `icustays`,

`services`, and `transfers` tables. A general suggestion to get a high level process model from all events in a group of patients would be to combine activity level of administrative events and middle or high level of clinical events.

4) *Aggregating events*. There were a small number of activities with the same timestamp. This need to be aggregated to get the correct sequence of events in the model.

The final event log was saved in a standard comma separated values (.csv) format file, which is readable in ProM. All codes and results of the transformation adjustments are available in the github repository for reusability purpose.

Loading, the final step in this stage was done by importing the file to ProM, converting it to the XES standard, and processing it to discover process models using the available algorithms/ plug-in modules. Unless otherwise stated, all plug-ins in ProM were applied with default parameter settings.

4.3. Stage 2: Create control flow model and connect event log

The next stage was to create control flow models and link these to the event log, using selected plugins in ProM. This study uses BPMN Miner as the main plugin for process model discovery. The discovered process models provided answers to some of the research questions specified before. The event log was filtered and adapted based on insight gained from the preliminary model (e.g. removing rare activities or outlier cases, focusing on specific activities, etc.).

The first research question was Q1 (*What are the most followed paths and what exceptional paths are followed?*). This could be answered by using event logs from three different levels. For example, to find the most followed admission and Intensive Care Unit (ICU) stay paths, records in the `admissions` and `icustays` tables can be used, which resulted in process model as shown in Fig. 2.

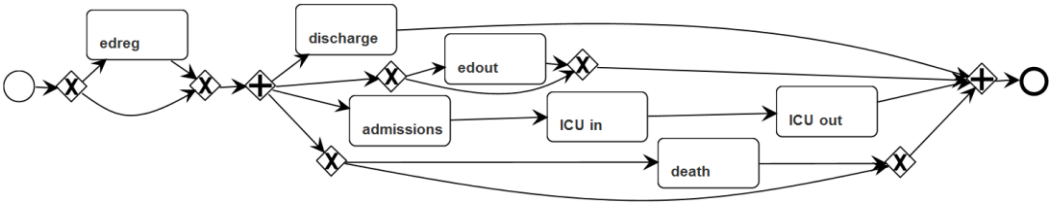

Fig. 2. Process model of admissions (fitness 0.971, precision 0.8808)

The process model in Fig. 2 was created with the "`BPMN Miner`" plug-in [23] in ProM, representing the 10,843 admission of 7,350 cancer patients having complete pathways. The evaluation was performed to the generated model with "`Replay a log for conformance checking`" plugin [24] in ProM. The generated model was evaluated using fitness and precision measures. Fitness quantifies the extent to which the discovered model can accurately reproduce the cases recorded in the log. Precision quantifies the fraction of the behaviour allowed by the model that is not seen in the event log. High fitness (0.971) and high precision (0.8808) means that the discovered model allows for the behaviour seen in the event log and does not allow for behaviour unrelated to what was seen in the event log.

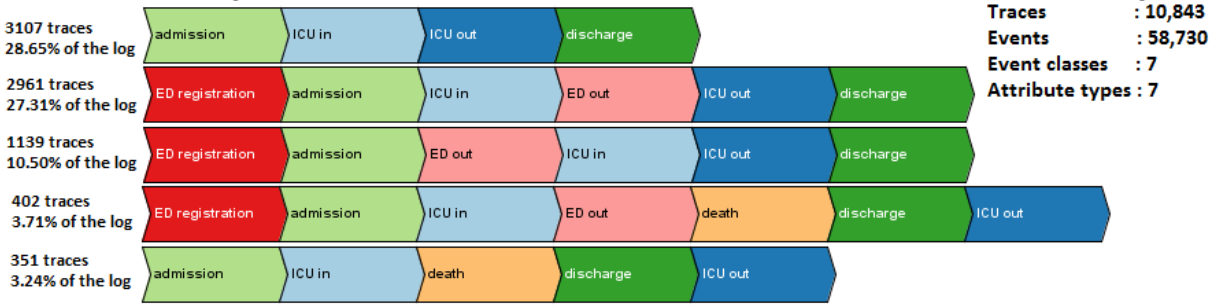

Fig. 3. The five most common trace variants representing 73.41% of traces

The most common trace variants are presented in Fig. 3. The pathways start with `admission` or `ED registration` and end with `discharge`. When a patient is admitted to the hospital, they could be admitted to a standard admission or registered through emergency department (ED). It can be seen that the most frequent activities are: `admission` and `discharge` activities. Other possible events are `ED registration`, `ED out` and `death`. We could also see that some variants are different because of different administrative steps, such as in the second and third variant (see Fig. 3). ICU in and ED out happened in the different order, while other events are the same, which indicate the different work order in cases where patients are moving from ED to the ICU.

Additional evaluation was done by 5-fold cross-validation. In this step, the original event log was randomly partitioned into 5 equal sized sub-eventlog. Of the 5 sub-eventlog, a single sub-eventlog was used as the validation data for testing the model, and the remaining 4 sub-eventlogs were used as training data. The cross-validation process was then repeated 5 times (the folds), with each of the 5 sub-eventlogs used exactly once as the validation data. The 5 results from the folds were then be averaged to produce a single estimation. From this step, the resulted fitness was 0.968644 and precision was 0.79726. The validation shows that both measures were expectedly lower than the conformance to itself, but both are still representing high values of fitness and precision.

Following the same method, we created process models from groups of patients with different cancer types to answer Q2 (*Are there differences in care paths followed by different patient groups?*). For example, the pathways of the three most common groups are presented in Fig. 4.

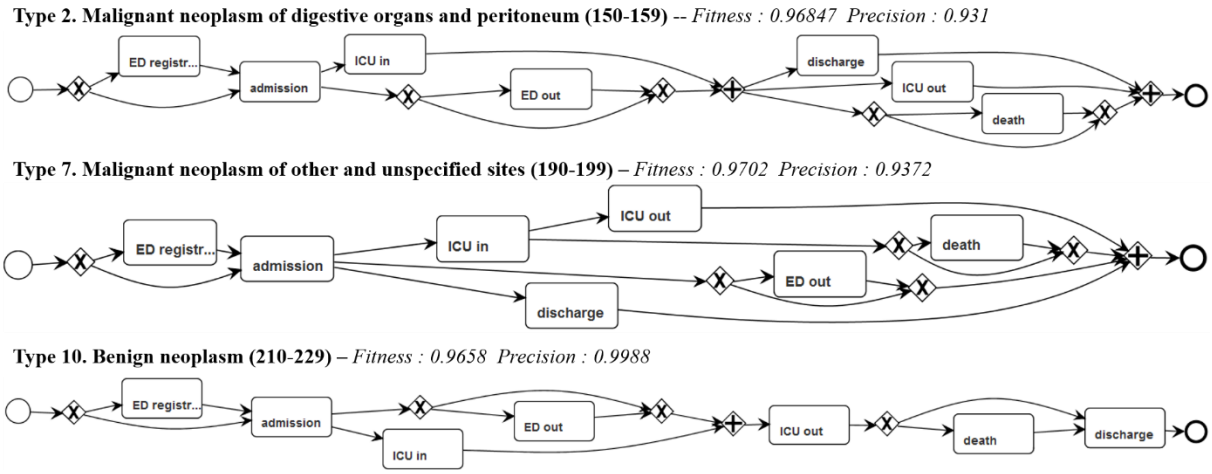

Fig. 4. Process models of three different cancer types

Our clinical domain expert reviewed the models based on their visual utility by comparing process models of cancer type 2, type 7, and type 10. The findings are: (1) `ICU in`, `ICU out`, `admission` and `discharge` are always happened in all three cancer types regardless of the sequence; (2) `ED registration`, `ED out` and `death` are possible events in all three types; (3) `admission` happened as the first event or after `ED registration` in all three types; (4) Some patients with cancer type 7 have been discharged after admission without other major events but type 2 and type 10 cancer patients have had more extensive care.

4.4. Stage 3: Create integrated process models

In stage 3, the models could be extended with other perspectives (e.g. date, time, and resources). In this study, the time information was used to analyse the waiting times for the admission pathways of all cancer patients. This was done to answer Q3 (*Where are the long waiting time activities in the process?*).

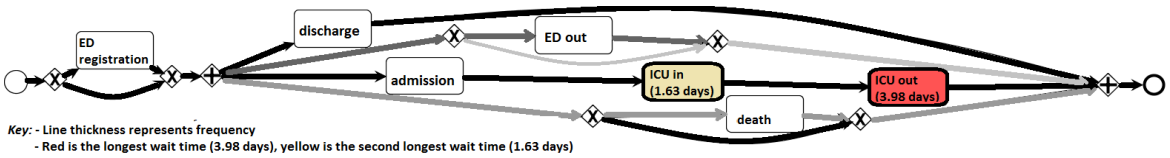

Fig. 5. Waiting time analysis (combined output from ProM plugins)

The BPMN in Fig. 2 was analysed with the "Replay a Log on Petri net for Performance/ Conformance Analysis" in ProM and the result was combined with the original BPMN, as shown in ***Fig. 5***. The analysis revealed that the longest waiting time is in ICU out (3.98 days in average), while the second longest waiting time is in ICU in (1.63 days in average). The long waiting times in ICU in and ICU out give an insight to dig deeper in the lower level of activities between ICU in and ICU out to understand which activities contribute to the long waiting time in ICU.

5. Discussion

The L* lifecycle was suitable for process mining in oncology using the MIMIC-III dataset, with the limitation that Stage 4 (Explicit operational support) was not applicable in this study due to the de-identified nature of the dataset. Stage 0 (Plan and justify) were performed by listing research questions based on frequently posed questions for process mining project in healthcare, which could be adapted in any clinical domains including oncology. All three questions listed have been answered through the experiments.

One important step during the Stage 1 (Extraction) was the transformation, which included the preprocessing of the dataset. Preprocessing should be based on a good understanding of the data and this study demonstrates how several preprocessing approaches were used to simplify the MIMICIII dataset. Preprocessing requires care, excluding a table might result in incomplete analysis and be insufficient to answer the research questions in the study.

Stage 2 (Creating control flow models) and Stage 3 (Create integrated process models) have been done in this study (Section 4) to answer some standard research questions. Process model comparisons were performed in a simple way in this study. Data quality issues were detected during these stages, and merit further analysis. This could be done by selecting events in the `allevents` table at three different levels.

6. Conclusion

This paper focuses on the applicability of process mining within healthcare domain, specifically cancer pathway, using the MIMIC-III dataset. MIMIC-III is representative of ICU and hospital datasets and its use can support reproducible research due to it being publicly available. This paper focused on gaining insights from the patient flow by implementing L* lifecycle using the control-flow and time perspective. It was shown that it is possible to mine complex hospital processes with existing techniques to discover and analyse process models.

Future work will explore several aspects. The first will be improving data quality with using data cleaning approaches. This will include the approaches to define a method to select events to include in the process discovery. The second area of focus will apply different algorithms for process mining to find process models with high conformance to the event log. While using only the Inductive Miner in this study, we recognise that there are many other algorithms available in ProM, where each of them are suitable for different characteristics of event log. This might lead to choosing the most suitable algorithm for this specific case study, or developing a new algorithm. Lastly, we will focus on advanced analysis with more clinical based research questions and comparison studies with other real-life datasets.

Acknowledgment

This research is supported by ClearPath Connected Cities Project. This is also a part of a PhD study funded by the Indonesia Endowment Fund for Education (LPDP).

References

[1] American Cancer Society, "The History of Cancer," 2011. [Online]. Available: www.cancer.net/patient/Advocacy and Policy/Treatment_Advances_Timeline.pdf. [Accessed: 09-Aug-2016].
[2] Cancer Research UK, "Your cancer type," 2014. [Online]. Available: http://www.cancerresearchuk.org/about-cancer/type/. [Accessed: 09-Aug-2016].
[3] W. M. P. van der Aalst, "Process Mining: Data Science in Action," in *Process Mining*, 2nd ed., Springer-Verlag Berlin Heidelberg, 2016, pp. 3–23.
[4] R. S. Mans, M. H. Schonenberg, M. Song, W. M. P. van der Aalst, and P. J. M. Bakker, "Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital," *Proc. BIOSTEC 2008*, vol. 25, pp. 425–438, 2008.
[5] R. Mans, W. M. P. van der Aalst, and R. Vanwersch, *Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes*, 1st ed. Springer International Pub, 2015.
[6] E. Rojas and J. Munoz-Gama, "Process mining in healthcare: A literature review," *J. Biomed. Inform.*, vol. 61, pp. 224–236, 2016.
[7] A. P. Kurniati, O. Johnson, D. Hogg, and G. Hall, "Process Mining in Oncology: a Literature Review," in *The 6th ICICM, IEEE*, 2016.
[8] B. W. Boudewijn van Dongen, Diogo R. Ferreira, "Business Process Intelligence Challenge (BPIC)," *IEEE Task Force on Process Mining*, 2011. [Online]. Available: http://www.win.tue.nl/bpi/doku.php?id=2011:challenge. [Accessed: 22-Apr-2016].
[9] R. Lenz and M. Reichert, "IT support for healthcare processes – premises, challenges, perspectives," *Data Knowl. Eng.*, vol. 61, no. 1, pp. 39–58, 2007.
[10] M. Poulymenopoulou, F. Malamateniou, and G. Vassilacopoulos, "Specifying Workflow Process Requirements for an Emergency Medical Service," *J. Med. Syst.*, vol. 27, no. 4, 2003.
[11] O. Johnson, P. S. Hall, and C. Hulme, "NETIMIS : Intelligent Simulation of Health Economics Outcomes using Big Data," *Pharmacoeconomics*, vol. 34, no. 2, pp. 107–114, 2015.
[12] K. HAYRINEN, K. SARANTO, and P. NYKANEN, "Definition, structure, content, use and impacts of electronic health records: A review of the research literature," *Int. J. Med. Inform.*, vol. 77, no. 5, pp. 291–304, May 2008.
[13] W. M. P. van der Aalst, *Process mining: discovery, conformance and enhancement of business processes*, 1st ed. Springer-Verlag Berlin Heidelberg, 2011.
[14] W. M. P. van der Aalst, "The application of Petri nets to workflow management," *J. circuits, Syst. Comput.*, vol. 8, no. 1, pp. 21–66, 1998.
[15] P. Wohed, W. M. P. van der Aalst, M. Dumas, A. H. M. ter Hofstede, and N. Russell, "Pattern-based Analysis of BPMN," *Lect. Notes Comput. Sci.*, vol. 4102/2006, pp. 161–176, 2006.
[16] W. M. P. van der Aalst, A. Adriansyah, A. K. A. de Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. van der Brand, R. Brandtjen, J. Buijs, A. Burattin, J. Carmona, M. Castellanos, J. Claes, and J. Cook, "Process Mining Manifesto," *BPM Work.*, vol. 99, pp. 169–194, 2011.
[17] MIT Laboratory for Computational Physiology, "MIMIC-III v1.3," 2015. [Online]. Available: https://mimic.physionet.org/about/releasenotes/. [Accessed: 03-Aug-2016].
[18] A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Data Descriptor : MIMIC-III , a freely accessible critical care database," *Sci. Data*, pp. 1–9, 2016.
[19] M. L. Van Eck, X. Lu, S. J. J. Leemans, and W. M. P. Van Der Aalst, "PM2: A process mining project methodology," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9097, pp. 297–313, 2015.
[20] S. J. J. Leemans, "Inductive visual Miner manual," pp. 1–16, 2017.
[21] 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," *LNCS (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7738 LNAI, pp. 140–153, 2013.
[22] "Online ICD9/ICD9CM codes - Neoplasms." [Online]. Available: http://icd9cm.chrisendres.com/index.php?action=child&recordid=1059. [Accessed: 16-Mar-2017].

[23] R. Conforti, M. Dumas, L. García-Bañuelos, and M. La Rosa, "BPMN Miner: Automated discovery of BPMN process models with hierarchical structure," *Inf. Syst.*, vol. 56, 2016.
[24] A. Adriansyah, "Replay a Log on Petri Net for Conformance Analysis plug-in.pdf | Algorithms | Logarithm."