Proceedings Paper:
Kusuma, G orcid.org/0000-0002-0208-125X, Kurniati, A orcid.org/0000-0002-4747-1067, McInerney, CD orcid.org/0000-0001-7620-7110 et al. (3 more authors) (2021) Process Mining of Disease Trajectories in MIMIC-III: A Case Study. In: Process Mining Workshops: ICPM 2020 International Workshops, Padua, Italy, October 5–8, 2020, Revised Selected Papers. 2nd International Conference on Process Mining (ICPM 2020), 04-09 Oct 2020, Padua, Italy (Online). Springer , pp. 305-316. ISBN 978-3-030-72692-8

https://doi.org/10.1007/978-3-030-72693-5_23

Process Mining of Disease Trajectories in MIMIC-III: A Case Study

Guntur Kusuma[1,3,*[0000-0002-0208-125X]], Angelina Kurniati[2[0000-0002-4747-1067]], Ciarán D. McInerney[1[0000-0001-7620-7110]], Marlous Hall[4[0000-0003-1246-2627]], Chris P. Gale[4[0000-0003-4732-382X]], Owen Johnson[1[0000-0003-3998-541X]]

[1] School of Computing, University of Leeds, UK LS2 9JT
[2] School of Computing, Telkom University, Bandung, Indonesia 40257
[3] School of Applied Science, Telkom University, Bandung, Indonesia 40257
[4] Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, UK LS2 9JT
*scgpk@leeds.ac.uk

Abstract. A temporal disease trajectory describes the sequence of diseases that a patient has experienced over time. Electronic health records (EHRs) that contain coded disease diagnoses can be mined to find common and unusual disease trajectories that have the potential to generate clinically valuable insights into the relationship between diseases. Disease trajectories are typically identified by a sequence of timestamped diagnostic codes very similar to the event logs of timestamped activities used in process mining, and we believe disease trajectory models can be produced using process mining tools and techniques. We explored this through a case study using sequences of timestamped diagnostic codes from the publicly available MIMIC-III database of de-identified EHR data. In this paper, we present an approach that recognised the unique nature of disease trajectory models based on sequenced pairs of diagnostic codes tested for directionality. To promote reuse, we developed a set of event log transformations that mine disease trajectories from an EHR using standard process mining tools. Our method was able to produce effective and clinically relevant disease trajectory models from MIMIC-III, and the method demonstrates the feasibility of applying process mining to disease trajectory modelling.

Keywords: Disease trajectories, Process mining, Electronic Health Records.

1 Introduction

There is a small but growing body of literature exploring the generation of disease trajectories using electronic health records (EHR) [1, 2]. The rich collection of patient data in the EHR is a valuable source to get an extensive trail of disease diagnoses over time [3]. Mining the trails of disease diagnoses and the temporal information may help to identify patterns in disease trajectories of clinical value. A better understanding of patterns of disease may advance precision medicine to improve care at an individual level [4] and improve medical understanding of common disease progression at the population level [5, 6]. A study by Jensen et al. [7] had identified the disease trajectories of a

large cohort by combining a data-driven and statistical approach. However, their trajectories were built based on overlapping pairs of diagnostic codes suggesting the presence of longer trajectories without confirming if such trajectories are available in the data. Based on this, we propose an improvement by incorporating process mining as a toolset and method for mining end-to-end disease trajectories.

Process mining utilises a set of tools to discover process models using data from an organisation's information system. Extracted data are transformed into an event log, a collection of activities and its corresponding timestamps, sometimes supplemented with additional attributes. There is now a large body of literature applying process mining to the domain of healthcare, typically focussed on discovery of actual care processes [8], conformance to guidelines and enhancement to improve the quality of healthcare services [9], the safety of the patients, and better management of resources [10, 11].

Jensen et al. [7] defined a disease trajectory as the patient's orderly series of diagnoses. The definition is comparable to the concept of a trace in process mining where a trace is the sequence of activities for an individual case [12]. We hypothesise that it should be feasible to apply process mining to discover a disease trajectory model [2]. To the best of our knowledge, this is the first time process mining has been used to identify disease trajectories from a real world EHR.

In this paper, we present a novel disease trajectory mining method using process mining techniques applied to the MIMIC-III open access EHR database. We identified the sequence of diagnoses (trace) based on the temporal aspect of the patients' admissions, broke down each trace into pairs of diagnoses, statistically analysed the pair's correlation and represented the identified disease trajectories using a directly-followed graph produced by standard process mining visualisation tools [12]. The research questions are as follow: *Q1-Can disease trajectories be identified using a process-mining approach? Q2-What are the most followed trajectories and what exceptional trajectories are followed? Q3-Are there differences in trajectories followed by different patient groups (by sex, by age group, by mortality status)?* And, *Q4-What are the longest and shortest average time transition trajectories?*

2 Background

Process mining provides a set of techniques and tools to uncover the real behaviour of processes from a range of perspectives including, but not limited to [12]: control-flow, performance, conformance, and organisational. There are three types of process mining: first, process discovery to generate process models from event log data, second, process conformance to check either a process model conforms to an event log or vice versa and third, process enhancement to improve a process model using the information of the actual process recorded in the event log [12].

In healthcare, process mining techniques may help the clinicians answer questions associated to each characteristic of the healthcare processes (e.g. primary care, secondary care, tertiary care, etc.) [8]. The rich information in the EHR is the source of answer

to the four types of data science questions: "*what happened?*", "*why did it happen?*", "*what will happen?*", and "*what is the best that may happen?*". In this study, we followed the most widely used methodology, the PM^2 framework, which describes six process mining stages and defines the set of activities to complete each stage.

The diagnostic codes available within electronic health records result from diagnostic decisions made by clinical specialists after considering the patient's health problem [13]. Jutel [14] described the diagnosis as a process of assessing and making a formal judgement based on a specific physical symptom that takes place at a particular time involving both patient and doctor. Once the disease is determined it is recorded in the EHR using standard diagnostic codes such as the World Health Organisation's International Classification of Diseases (ICD) [15].

3 Method

The goal of this case study was to identify patients' disease trajectories using a process-mining approach. We conducted a retrospective cohort study of patients who were admitted to critical care using the MIMIC-III database as our data source [16]. The MIMIC-III database contains a detailed record of patients' clinical care that has been de-identified to respect the sensitive nature of the data. It is available online to researchers (https://mimic.physionet.org) under an open access policy. We obtained access through two mandatory steps: a training program in human research subject protections and a data user agreement. The Process Mining Project Methodology (PM^2) was followed in this study as the methodology allows us to have multiple research questions that require iterations of analyses [17].

3.1 Data source for the case study

MIMIC-III provides a database of de-identified electronic health records containing the medical history from 2001 to 2012 of 46,520 critical care patients extracted from the EHR of the Beth Israel Deaconess Medical Centre in Boston, USA [16]. The database includes data on patient demographics, laboratory tests, diagnostic codes (in ICD-9 coding standard), medications, bedside monitoring, clinicians' notes and reports, and death records (linked to Social Security Death Index for outpatient death). As part of the anonymisation process, the timestamps used in the MIMIC-III dataset have been intentionally shifted into the future (between 2100 and 2200) by a random offset generated for each patient. This means that the sequence of disease codes and the time intervals between disease codes has been preserved for individual patients but no comparisons between patients are possible. This does not affect disease trajectory mining, but does limit other process-mining approaches such as the identification of bottlenecks. Our group has experience of applying process mining to MIMIC-III and in earlier work have published a data quality assessment on the suitability of the various MIMIC-III data components that are compatible with process mining [18].

3.2 PM² for Disease Trajectory Mining

In this section, we identify those sections of the PM² that we have adapted for disease trajectory mining. For a full understanding of the PM² method see [17].

In Stage 1 (Planning), our research questions were identified from a literature review and confirmed by a project team composed of a clinician, and epidemiologist and process mining and data science researchers.

In Stage 2 (Extraction), we defined the scope by determining the granularity level of data, the time period, and attributes of interest. The MIMIC-III database contains admissions of adult patients aged 16 years old or older [16] who were admitted to the hospital between 1 June 2001 and 10 October 2012. Only patients with at least two admissions were selected to capture the progression of the disease. Patients were followed up for mortality status until the last available discharge as the last censoring date and time for those who died within the hospital. The censoring date for patients who died outside of the hospital is the date recorded in the social security master death index in the MIMIC-III database. We used the first 3-digit ICD-9 codes to indicate diagnoses, [19] but excluded codes known not to be related to development of diseases, e.g. administration codes. Event data were extracted from the ADMISSIONS, PATIENTS, and DIAGNOSES_ICD tables in MIMIC-III database as the input for creating an event log (**Table 1**). The time of admission was used as the activity timestamp and the diagnostic code as the activity name. The patients were grouped according to their age in bands of 5 years. The attribute of age group was calculated from the patient's age at first admission.

In Stage 3 (Data Processing), we created the event log as defined in the PM² by creating the views, then filtering and enriching them. The case identifier for each event was taken from the patient identifier (`subject_id`), the diagnostic code was used as the event name (`diagnosis_code`), and the admission time as the timestamp (`admittime`). The event log was filtered by removing recurring diagnostic codes (retaining the first occurrence), then reapplying the exclusion of patients with only one diagnostic code. The sequences of diagnostic codes for each patient in the event log informed a set of ordered pairs of diagnostic codes, D1→D2, where the diagnostic code D1 preceded the diagnostic code D2. For example, a patient's event log, D1→D2→D3, informed two ordered pairs of diagnostic codes, D1→D2 and D2→D3. We excluded ordered pairs that occurred only once. To measure the strength of association between the ordered pairs, we compared the probability of diagnosis D2 occurring among patients who did and did not have a D1 diagnosis previously in the event log. This relative risk (RR) [20] indicated whether the D2 diagnosis was more incident in the group with a D1 diagnosis (RR > 1), less incident in the group with a D1 diagnosis (RR < 1), or equivalent (RR = 1). The RR is calculated as

$$RR = \frac{(a/(a+b))}{(c/(c+d))} \tag{1}$$

where a is the number of patients having D1 and D2, b is the number of patients having D1 but not D2, c is the number of patients without having D1 but having D2, and d is the number of patients neither having D1 nor D2.

Table 1. Source of the required data from MIMIC-III database

Variables	Table source in MIMIC-III	Field name
Case identifier	PATIENTS	subject_id
Event	DIAGNOSES_ICD	hadm_id, icd9_code, seq_num
Activity name	DIAGNOSES_ICD	icd9_code (first 3 digits)
	ADMISSIONS	hospital_expire_flag
	PATIENTS	expire_flag (translated into 1:Dead, 0:End of data)
Time stamps	ADMISSIONS	admittime,dischtime, deathtime
	PATIENTS	dod, dod_hosp, dod_ssn,
Sex	PATIENTS	gender
Age*	PATIENTS	dob
	ADMISSIONS	admittime
Age group**	PATIENTS	dob
	ADMISSIONS	admittime

* the age calculation using PATIENT's dob and ADMISSIONS's admittime.

** the variable was added to group the patients' age.

Following Jensen et al [7], only pairs with RR > 1 were carried forward for further processing. For a given pair of diagnoses D1 and D2, it was possible for both D1→D2 and D2→D1 trajectories to satisfy the RR > 1 threshold. Our goal was to identify disease trajectories that were acyclic, so we carried forward the dominant directionality of a given pair of diagnostic codes, only. We applied one-tailed binomial tests [21] to define the dominant directionality of pairs, i.e. D1→D2 or D2→D1. Using a significance level of $\alpha = 0.05$, only ordered pairs of diagnostic codes with one statistically significant direction were carried forward to define the final pairlog.

The final pairlog was transformed back into an event log and recurring diagnoses in each trace were merged to avoid loops. The event log was then enriched by adding attributes of age at admission, sex, age group and the mortality status. These attributes were not used to define the disease trajectory models, but allowed post-hoc analyses to determine differences between disease trajectories according to each attribute. The enriched event log was then loaded into ProM, an open-source process mining tool (https://promtools.org). A START and END event was added to every case in the event log to provide common start and end points of traces. The final event log then converted into the XES format. Common traces were grouped in trace variants using the Explore Event Log (Trace Variants/ Searchable/ Sortable) feature in ProM [22].

In Stage 4 (Mining and Analysis) we used ProM to analysed the event log to identify unique trace variants, performed process discovery, visualised the discovered model and performed conformance checking. For process analysis, we calculated descriptive summary statistics of the disease trajectories that were identified, including stratification by patient groups. The event log was visualised using the Explore Event Log (Trace variants/ Searchable/ Sortable). The Interactive Data-aware Heuristics Miner (iDHM) [23] plug-in was used to discover the disease process models.

The quality of the discovered models were evaluated using replay fitness, precision and generalisation [24]. Replay fitness is a measure of how many traces from the log can be reproduced in the process model, with penalties for skips and insertions. Precision is a measure of how 'lean' the model is at representing traces from the log. Lower

values indicate superfluous structure in the model. Generalisation is a measure of generalisability as indicated by the redundancy of nodes in the model; The more redundant the nodes, the more variety of possible traces that can be represented. The value of each measure represents by a number between 0–1. Discovery and conformance checking used plugins in ProM. The Replay a Log on Petri Net for Conformance Analysis plug-in for measuring the fitness [25], Align-ETConformance plug-in [26] for the precision, and the Measure Precision/Generalization plugin for measuring the generalisation. Other tools used in this study were PostgreSQL as the database management system of MIMIC-III, and Python through Jupyter Notebook [27].

4 Results

An event log was extracted from an EHR to identify disease trajectories, pairs of diagnoses were identified and analysed for correlation measurement and tested for directionality. The discovery algorithm is applied to produce the disease trajectory model and represented using the directly-followed graph.

In Stage 1 (Planning), we aimed to mine the disease trajectory agnostically without any specific selection of diagnosis and time window. Following the literature review in section 2, we defined the main research question as: (*Q1) Can disease trajectories be identified using a process-mining approach?* Further questions added which were motivated by the frequently posed question for process mining in healthcare [28]: *(Q2) What are the most followed trajectories and what exceptional trajectories are followed?(Q3) Are there differences in trajectories followed by different patient groups (by sex, by age group, by mortality status)? (Q4) What are the longest and shortest average time transition trajectories?*

In Stage 2 (Extraction), Of the 58,976 unique admissions in MIMIC-III from 46,520 patients, there were 6,984 unique ICD-9 diagnostic codes used for 651,000 diagnoses. From this dataset, we excluded 172,685 (26.5%) diagnostic codes that are medically known to be codes related to external factors not directly related to the development of diseases [5], including pregnancy (ICD-9 3-digit codes 630-679, 760-779), general symptoms and signs not related to a disease (780-799), external cause (800-999, E800-E999), and administration (V01-V89). We further excluded 436,483 (67%) secondary diagnostic codes and focused on the 41,832 primary diagnostic codes whilst there will be valuable opportunity in exploring the secondary diagnostic codes.

In Stage 3 (Data Analysis), we composed the selected variables in a way that follows the minimum requirements of event log (see **Fig. 1.a**). The traces of each patients are illustrated in **Fig. 1.b**. We removed 2,692 (16.2%) recurrent diagnoses, retained the first occurrence, excluded patients with only one admission, and subsequently excluded patients who were less than 16 years old at their first ever admission. A total of 4,911 patients remained in the event log consisting of 11,725 diagnostic codes. **Fig. 1** shows the transformation of event logs into a log of ordered pairs of diagnostic codes (pairlog)(see **Fig. 1.c**). The resulting pairlog contained 6,814 ordered pairs of diagnostic codes. Only 3,781 pairs remained after filtering for RR > 1 and the binomial tests for

directionality suggested there were 826 ordered pairs of diagnostic codes with a statistically significant dominant direction. The resulting data contained 796 traces where each trace represents a patient's disease trajectory.

subject_id	diagnostic_code	timestamp
21	410	11/09/2134 12:17
21	038	30/01/2135 20:50
124	433	24/06/2160 21:25
124	441	17/12/2161 03:39
124	440	21/05/2165 21:02
124	569	31/12/2165 18:55

(a) The extracted event log

#21: 410→038

#124: 433→441→440→569

(b) The trace of diagnosis

subject_id	Antecedent	Subsequent	Time1	Time2
21	410	038	11/09/2134 12:17	30/01/2135 20:50
124	433	441	24/06/2160 21:25	17/12/2161 03:39
124	441	440	17/12/2161 03:39	21/05/2165 21:02
124	440	569	21/05/2165 21:02	31/12/2165 18:55

(c) The *pairlog*

Fig. 1. Illustration of the transformation steps of event log for pairwise analysis. (a) The extracted event log from MIMIC-III; (b) the illustration of traces of diagnoses for each patient; (c) the transformed event log into pairlog.

In the last step of filtering, we transformed the pairlog back to an event log and enriched with age at admission, sex, age group and the mortality status. We then loaded the enriched event log into ProM, artificial 'START' and 'END' events were added and then analysed the trace variants using the Explore Event Log feature. Among the 796 traces, we further removed twenty traces that were unique to a single, individual patients as part of good anonymisation practice. Finally, the 776 common traces found in the event log were grouped into 81 trace variants.

In Stage 4 (Mining and Analysis), there were eighty one unique trace variants informed the processing discovery algorithms to answer the Q1. The conformance of the discovered disease trajectory model demonstrated fitness = 0.93, precision = 0.94, and generalisation = 0.92. Further evaluation was done by 5-folds cross-validation where the original event log was randomly divided into five groups of sub-event log equally. One sub-event log was used as the validation data and the remaining four sub-event logs as training data. The cross-validation process was done five times to allow each sub-event log used once as the validation data. The average value from the cross-validation are expected to be lower than the conformance, resulting fitness = 0.92 (SD: 0.006), precision = 0.82 (SD: 0.06), and generalisation = 0.88 (SD: 0.02). This suggests that the discovered trajectory model (**Fig. 2**) is robust to sampling, allows the traces seen in the event log, is precise enough to not allow behaviour unrelated to what was seen in the event log, and general enough to reproduce future behaviour of the trajectories.

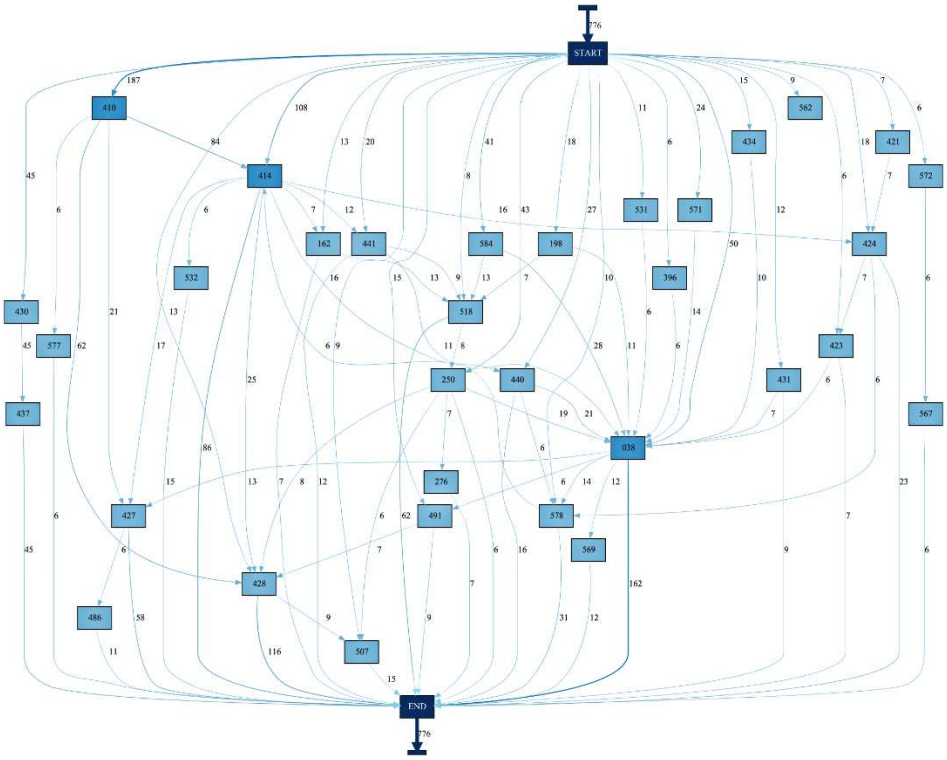

Fig. 2. The directly-follow graph representation of Disease Trajectory Model of Critical Care patients in MIMIC-III with the minimum case frequency = 6.

In respond to the Q2, among 776 patients there are 81 distinct trajectories (**Table 2**). The most-followed trajectory (n=80; 10.3%) was acute myocardial infarction to ischemic heart disease, which is consistent with the published literature [7, 29, 30]. Septicaemia occurred most frequently (n=212; 27.3%), both as a precedent (n=50; 6.4%) and subsequent (n=162; 20.9%), with mortality in the end (n=143; 66.9%). This supported previous findings that it is associated with morbidity and mortality [16, 31]. There are three exceptional trajectories of two patients each (0.26%) (**Table 2**).

Table 2. The three most-common and least-common trace variants.

Traces (%)	Trace Variant	Median (months)	Dead (%)	Male (%)
80 (10.31%)	START→410→414→END	6.5	75	70
62 (7.99%)	START→410→428→END	3.9	72.58	54.84
45 (5.80%)	START→430→437→END	3.9	4.44	35.56
…	…	…	…	…
2 (0.26%)	START→410→427→486→END	28.3	100	50
2 (0.26%)	START→507→491→482→END	43.6	50	100
2 (0.26%)	START→518→250→038→END	14.6	100	0

ICD-9 Codes translation: 038 = Septicaemia, 250 = Diabetes mellitus, 410 = Acute myocardial infarction, 414 = Ischemic heart disease, 427 = Cardiac dysrhythmias, 428 = Heart failure, 430 = Subarachnoid haemorrhage, 437 = Other and ill-defined cerebrovascular disease, 482 = Other bacterial pneumonia, 486 = Pneumonia, organism unspecified, 491 = Chronic bronchitis, 507 = Pneumonitis due to solids and liquids, 518 = Other diseases of lung.

The third question was (Q3) Are there differences in trajectories followed by different patient group? We answered the question by comparing trajectories by sex (male, female) and age band (18-34 years, 35-64 years, and >64 years). The male cohort consisted of 447 patients with the median duration of follow-up 6.98 months (IQR 1.6 – 28.2) where 252 cases (56.3%) ending in death. The most-common trajectory was acute myocardial infarction followed by other forms of chronic ischemic heart disease (56 cases, 12.5%) with median interval 6.5 months (IQR 1.5 – 35.3). In the female cohort, there were 329 patients with the median duration of follow-up 7 months (IQR 2 – 24.4) where 176 cases (54.4%) ending in death. The most-common trajectory was subarachnoid haemorrhage followed by other and ill-defined cerebrovascular disease (29 cases, 8.8%) with median interval 3.4 months (IQR 2.3 – 7.5). The most-followed trajectory in a group of 18 to 34-year-old cohort was diabetes followed by hypertensive chronic kidney disease (3 cases) with median interval 55.8 months (IQR 33 – 56.5). For the group of 35 to 64 years, there were 44 cases (14.5%) with acute myocardial infarction followed by ischemic heart disease, with median interval 7.8 months (IQR 1.9 – 39.7). Among 329 cases in this age group, there were 133 cases (40.4%) ending in death. Patients in >64 years, there were 293 (68.1%) deaths while the most-common trajectory was acute myocardial infarction followed by heart failure, with median interval 4.7 months (IQR 1.5 – 21.8).

The fourth question was (Q4) What are the longest and shortest average time transition trajectories? The longest disease progression at 63 months was *Ischemic heart disease* to *Diverticula of intestine* while the shortest progression was *Gastrointestinal hemorrhage* to *Liver abscess and sequelae of chronic liver disease* with average time transition is less than a month (0.98) (**Table 3**).

Table 3. The three longest and shortest average time interval trajectories in MIMIC-III.

Antecedent	Subsequent	Mean*	Median (IQR)**
A. The three longest average time interval trajectories (descending)			
Chronic ischemic heart disease	Diverticula of intestine	63	75.9 (54 – 84.8)
Chronic ischemic heart disease	Occlusion of cerebral arteries	52.7	51.2 (40.4 – 52.6)
Chronic ischemic heart disease	Heart failure	46	41.5 (4.6 – 89.7)
B. The three shortest average time interval trajectories (ascending)			
Gastrointestinal hemorrhage	Liver abscess and sequelae of chronic liver disease	0.98	0.81 (0.6 – 1.3)
Other diseases of endocardium	Other diseases of pericardium	1	0.8 (0.6 – 1.13)
Chronic bronchitis	Other bacterial pneumonia	2.2	2.2 (1.6 – 2.7)

*Mean is in months. **Median is in months (IQR); IQR = interquartile range.

5 Discussion

We present a case study of 776 patient admissions associated with 81 different disease transitions to demonstrate the feasibility of using a process-mining approach to reveal disease trajectories using a hospital electronic health record database. We show that the PM^2 framework is suitable for mining disease trajectories and is complemented by the addition of descriptive summary statistics in Stage-3 (Data Processing). Our approach

applies a number of transformations to the data, which were adapted from published disease trajectory methods for constructing selected pairs of diagnoses with strong correlation, followed by testing the pairs' directionality to form the trajectories.

Process mining offers techniques to discover disease trajectories and measure the quality of the algorithm to discover the trajectory model. In this work we presented replay fitness, precision, generalisation and cross-validation to validate the model. The process-mining approach opens opportunities to cross-reference discovered disease trajectories with other critical care event data by defining workflows that can actioned using widely-available software. By conducting conformance checking, we have the indicators to show if the discovered model has a good quality. We note that the earlier study by Jensen et al. [7], did measure the robustness of their discovered disease trajectory model with one indicator that is similar to the replay fitness in process mining. This approach is useful to validate that the final model conforms closely to the data.

A particular benefit of the process-mining approach to constructing disease trajectories is that it may provide summaries of cases, events and time interval between occurrences of disease. For example, our method identified the trajectory of *acute kidney injury* (AKI) (584) followed by *septicaemia* (038) with an average interval of 16.22 months. This finding supports the conclusion of [32] where sepsis was a frequent consequence after AKI in intensive care setting. Also, the process-mining approach could provide an estimation of sepsis development after AKI as suggested in [33]. Our method also incorporates additional case attributes that easily facilitate outputs to be stratified by specific characteristics, e.g. sex, age group, and mortality status. For example, although the data were not pre-stratified for females, process mining tools made it easy to query the event log to reveal a dominant trajectory in females – *subarachnoid haemorrhage* (430) followed by *other and ill-defined cerebrovascular disease* (437) – that agrees with previous research [34].

6 Conclusion

In this paper, we have presented the mining of disease trajectories using a process-mining approach. The mining used the MIMIC-III dataset which is comparable to many databases from EHR systems in use at hospitals across the world. Our study included the use of PM^2 framework to mine a representative disease trajectory model from an EHR and addressed quality dimension standards. This study opens opportunities for future works in implementation of the technique using population sized EHR data. We believe the association of pairs of diagnoses might be improved by null hypothesis significance testing of relative risk rather than magnitude-based testing. Future work might assess the sensitivity of the method to the choice of process discovery algorithm used to mine the disease trajectory model.

7 Acknowledgements

The research was supported by the National Institute for Health Research (NIHR) Yorkshire and Humber Patient Safety Translational Research Centre (NIHR YH PSTRC) and the Indonesia Endowment Fund for Education (LPDP).

References

1. Allam, A., et al., *Patient Similarity Analysis with Longitudinal Health Data.* arXiv preprint arXiv:2005.06630, 2020.
2. Kusuma, G., et al. *Process Mining of Disease Trajectories: A Feasibility Study*. in *13th International Conference on Health Informatics*. 2020.
3. Weber, G.M., K.D. Mandl, and I.S. Kohane, *Finding the missing link for big biomedical data.* JAMA, 2014. **311**(24).
4. Jensen, P.B., L.J. Jensen, and S. Brunak, *Mining electronic health records: Towards better research applications and clinical care*. 2012. p. 395-405.
5. Hanauer, D.A. and N. Ramakrishnan, *Modeling temporal relationships in large scale clinical associations.* J Am Med Inform Assoc, 2013. **20**(2): p. 332-341.
6. Rothman, K.J. and S. Greenland, *Causation and causal inference in epidemiology.* American Journal of Public Health, 2005. **95**(SUPPL. 1): p. S144-50.
7. Jensen, A.B., et al., *Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients.* Nature Comm., 2014. **5**(May): p. 1-10.
8. Mans, R.S., W.M.P. van der Aalst, and R.J.B. Vanwersch, *Process Mining in Healthcare Evaluating and Exploiting Operational Healthcare Processes*. 1 ed. 2015: Springer International Publishing. 91.
9. Partington, A., et al., *Process Mining for Clinical Processes: A Comparative Analysis of Four Australian Hospitals.* ACM Trans. Manag. Inform. Syst. Article, 2015. **5**(19).
10. Rojas, E., et al., *Process mining in healthcare: A literature review.* Journal of Biomedical Informatics, 2016. **61**(April): p. 224-236.
11. Fernandez-Llatas, C., et al., *Process Mining Methodology for Health Process Tracking Using Real-Time Indoor Location Systems.* Sensors, 2015. **15**(12): p. 29821-29840.
12. van der Aalst, W.M.P., *Process Mining: Data Science in Action*. 2 ed. 2016: Springer-Verlag Berlin Heidelberg. 467.
13. National Academies of Sciences, Engineering and Medicine, *Improving Diagnosis in Health Care*. 2015, Washington, DC: The National Academies Press.
14. Jutel, A., *Sociology of diagnosis: a preliminary review.* Sociology of Health & Illness, 2009. **31**(2): p. 278-299.
15. World Health Organization, *Classification of Diseases*. 2019.
16. Johnson, A.E.W., et al., *MIMIC-III, a freely accessible critical care database.* Scientific Data, 2016. **3**: p. 160035-160035.
17. van Eck, M.L., et al., *PM2: A process mining project methodology*, J. Zdravkovic, M. Kirikova, and P. Johannesson, Editors. 2015, Springer, Cham. p. 297-313.
18. Kurniati, A.P., et al., *The assessment of data quality issues for process mining in healthcare using MIMIC-III , a publicly available e-health record database*. 2017.
19. National Center for Health Statistics, *ICD-9-CM Official Guidelines for Coding and Reporting*. 2011.

20. Tenny S, H.M. *Relative Risk*. 10 July 2020; Available from: https://www.ncbi.nlm.nih.gov/books/NBK430824/.
21. Kang, S.-H. and C.W. Ahn, *Tests for the homogeneity of two binomial proportions in extremely unbalanced 2 x 2 contingency tables.* Statistics in medicine, 2008. **27**(14): p. 2524-2535.
22. Mannhardt, F., *Tools & Software — ProM — Event Log Explorer*. 2018.
23. Mannhardt, F., M. De Leoni, and H.A. Reijers. *Heuristic mining revamped: An interactive, data-Aware, and conformance-Aware miner*. in *BPM 2017*. 2017. CEUR-WS.org.
24. Buijs, J.C.A.M., B.F. Van Dongen, and W.M.P. Van Der Aalst, *On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery.* 2012.
25. Ardiansyah, A., *Replay a Log on Petri Net for Conformance Analysis-plugin.pdf*. 2012.
26. Adriansyah, A., et al., *Measuring precision of modeled behavior.* Information Systems and e-Business Management, 2015. **13**(1): p. 37-67.
27. Kluyver, T., et al., *Jupyter Notebooks—a publishing format for reproducible computational workflows*. 2016. 87-90.
28. Mans, R.S., et al. *Process Mining in Healthcare: Data Challenges when Answering Frequently Posed Questions*. 2012. Berlin, Heidelberg: Springer.
29. Asaria, P., et al., *Acute myocardial infarction hospital admissions and deaths in England: a national follow-back and follow-forward record-linkage study.* The Lancet Public Health, 2017. **2**(4): p. e191-e201.
30. Hall, M., et al., *Multimorbidity and survival for patients with acute myocardial infarction in England and Wales: Latent class analysis of a nationwide population-based cohort.* PLoS Medicine, 2018. **15**(3).
31. Sakr, Y., et al., *Sepsis in Intensive Care Unit Patients: Worldwide Data From the Intensive Care over Nations Audit.* Open forum infectious diseases, 2018. **5**(12): p. ofy313-ofy313.
32. Mehta, R.L., et al., *Sepsis as a cause and consequence of acute kidney injury: Program to Improve Care in Acute Renal Disease.* Intensive care medicine, 2011. **37**(2): p. 241-248.
33. Peerapornratana, S., et al., *Acute kidney injury from sepsis: current concepts, epidemiology, pathophysiology, prevention and treatment.* Kidney international, 2019. **96**(5): p. 1083-1099.
34. Eden, S.V., et al., *Gender and ethnic differences in subarachnoid hemorrhage.* Neurology, 2008. **71**(10): p. 731-735.