A PROCESS MODELING STRATEGY TO LEARN ISCHEMIC STROKE TREATMENT PATTERNS FROM ELECTRONIC MEDICAL RECORDS

By

Lina Sulieman

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

August, 2014

Nashville, Tennessee

Approved

Bradley Malin, Ph.D., Chair

Nancy Lorenzi, Ph.D.

Jeremy Warner, MD, MS.

Daniel Fabbri, Ph.D.

ACKNOWLEDGEMENTS

First, I would like to sincerely thank my advisors, Dr. Bradley Malin and Dr. Daniel Fabbri, for their tremendous mentoring, and guidance throughout this work. I am also grateful to the other members of my committee, Dr. Nancy Lorenzi, Dr. Jeremy Warner, and Dr. Josh Peterson for providing guidance in some parts of the work.

I would also like to thank the Department of Biomedical Informatics including faculty, staff and students, and Steve Nyemba for his assistance in providing access to the data and tools I needed during my project.

Finally I would like to thank my family, and especially my mother, for supporting me, and my Fulbright grant for giving me the chance to peruse this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

There are various studies which indicate that the adoption of health information technology (HIT) can improve patient outcomes and reduce medical errors [8]. Moreover, implementing electronic medical record (EMR) systems can benefit clinicians by providing information about a patient in real time and assisting in clinical decision making [1,8,25]. The EMR is integral to all aspects of a patient's interaction with the healthcare enterprise. From a patient's admission until discharge, healthcare providers interact with a patient's EMR and apply necessary clinical actions to treat the patients. The patterns by which the healthcare provider interacts with the EMR system can identify the patient-specific information needs or patient-centered workflow [11].

In general, clinical workflows are patterns of actions that healthcare providers apply to accomplish tasks associated with a patient's treatment [8]. Standardizing the clinical workflow by implementing guidelines and protocols can reduce the variability in the treatment process and ensure the effective utilization of EMR [43,44]. To share and standardize the implementation of clinical guidelines, the InterMed Collaboratory (which is a collaborative partnership among investigators from Columbia, Harvard, and Stanford universities) worked to develop shared infrastructural software, tools, and system components. The InterMed collaborators developed GuideLine Interchange Format (GLIF) to encode the guideline as computer-interpretable guidelines which is more relevant way instead of the referring to all published text guideline [44,46]. An evidence-based consensus has to be developed whether manually reviewing the medical literature or automatically discovering the current clinical workflow using data mining techniques [44]. Implementing GLIF reduces the variability in translating, sharing and implementing the standard guidelines [46], however, the treatment of patients with the same diagnosis may differ due to multiple factors, such as the healthcare team who treats the patient, the demographics and comorbities associated with the patient, and the laboratory test results [11,23]. Understanding the clinical workflow and treatment patterns may identify

bottlenecks in treatments processes, evaluate the current treatment plans, and compare implementation across clinics and within the same clinic.

Over the past several decades, workflow mining techniques have been developed and applied to understand the current processes in place at an organization and, subsequently, to improve its performance (e.g. [31,48,56,58]). In particular, a collection of workflow mining methodologies, have been designed to utilize the log of information system events, analyze the outcome of the existing system, and infer the followed processes [8,15,56]. Process mining relies on ordered events stored in the system logs to discover the current process, find common performed steps, and compare the observed processes to a desired process flow [31].

To assess the difference in quality of care, using inferred process measures may require less data collection (e.g., lower number of cases required) in comparison to traditional outcome quality assessment strategies such as mortality [34,35]. Process mining techniques have been invoked to model and investigate the processes in various industrial settings such as supply chains, banks (e.g., opening accounts), government agencies [54]. Translating such techniques into the clinical setting may provide intuition into the pathways patients follow. Moreover, the common steps in such pathways may support evaluation of the current clinical practices and improve the treatment processes by reducing uncertainty and achieving treatment goals within the required treatment timeline [22,31,32,43]. Standardizing treatment processes may further reduce variability in the data collection and enable discovery of the ideal point at which a healthcare facility should integrate decision support tools or present information to users of the EMR [43].

However, there are a number of challenges to the application of existing process mining techniques for healthcare data. First, the healthcare environment is highly dynamic [30], while, the clinical processes are complex and mutli-disciplinary. There are many factors that may influence the treatment process (both at the clinical and organizational level), such as i) the care providers who are involved in the treatment process, ii) the clinical protocol that may affect the treatment process, and iii) new regulations or policies that require organizational changes [48,49]. Second, healthcare organizations that rely upon

EMRs are complex cyberphysical systems that rely heavily on human factors that perform a variety of actions that are based on various factors outside of the digital domain. As a consequence, the variability between patients within the same cohort or treated for the same reason can be high. Third, healthcare providers may treat the same condition in different ways depending on the social characteristics (e.g., primary care physician), economic characteristics (e.g., patient insurance), and patient's health status [2,50].

Most of the processing mining techniques used petri net to describe the process model in a graphical model. Nevertheless, recent studies have applied multiple sequence alignment (MSA) to infer the common behaviors associated with performing certain processes [7,22]. To date, most studies that have applied process mining in the clinical domain have assumed systems that 1) are devoid of noise and 2) exhibit only a small variation in the order or number of steps in the treatment plan for patients with the same diagnosis. Moreover, existing algorithms focus on an organizational view for the treatment process and fail to address the clinical process that affects the role of the person who will be involved in the treatment or how sudden clinical events can change a treatment plan. As such, direct application of traditional process mining methods is unlikely to provide the most common steps that are invoked to treat a specific cohort of patients. Evidence already suggests that doing so will provide a very complex and spaghetti-like model [5,7,22,35].

This thesis aims to overcome the aforementioned limitations in process modeling for clinical systems. In particular, this thesis makes several specific contributions.

- **Clinical Process Mining**: We introduce a system, called Treatment Mining using Frequent Sequential Patterns (TMFSP). This is a multi-step learning approach to detect frequent treatment patterns from the standard actions that healthcare providers perform to treat patients admitted for a specific diagnosis. First, TMFSP forms the treatment sequence from the clinical events documented in a patient's EMR, such as medication and laboratory order sets. Second, it uses the frequent patterns to reduce the dimension of the data and remove the noise. Third, it

applies MSA to discover the shared frequent subsequence patterns for patients in a cohort and represent the sequences using the common sequential patterns.

- **TMFSP Evaluation:** We evaluated our methodology by mining the clinical process for a cohort of 133 patients diagnosed with ischemic stroke patients over 4 months. The results illustrate that the patients' treatment plans include 2,020 patterns that consists of 7 medications and 12 laboratory tests. Moreover, TMFSP generated a common clinical pathway that the patients' treatment sequences share. In addition, it was discovered that Insulin and Beta Blockers were excluded from a subpopulation treatment due to lipid metabolism disorders influence.

The reminder of the thesis is organized as follows. In Chapter 2, we review related research in process mining, with a particular focus on workflows. In Chapter 3, we present the TMFSP approach for modeling patient treatment processes. Chapter 4 introduces the cohort of ischemic stroke patient records extracted from the EMR of the Vanderbilt University Medical Center and the series of experiments applied to evaluate TMFSP. In Chapter 5, we discuss the main experimental findings and limitations of the study. Finally, Chapter 6 summarizes the next steps and logical extensions to this work.

CHAPTER 2

BACKGROUND AND RELATED WORK

A clinical pathway is a standardized treatment pattern that implements guidelines and protocols formed by clinical experts [21,38]. The standard pathway implies that specific interactions with the EMR system occur at a defined time or in a known order. Some studies implemented qualitative methods, such as surveys or observational studies, to collect data from EMR users to define clinical workflow [11]. During the past years, different algorithms and models were developed to perform process mining in the industrial domain and healthcare domains. It will be instructive to describe some of the algorithms and techniques that have been used, examine their success, highlight their shortcomings, and learned lessons that influenced our model. In the first section, we describe the usage pattern of an EMR system and its relation with clinical processes. In the second section, we discuss the usage of petri nets and the limitation of this framework on mining healthcare processes. The third section discusses multiple sequence alignment (MSA) and its usage in the domain of process mining. The fourth section describes the data preprocessing that can be applied on the clinical data to reduce the data dimension and remove noise.

2.1 EMR Utilization

The availability of patient information plays a crucial role in increasing the EMR adoption [53]. When a patient is hospitalized, healthcare providers use and integrate different clinical data types to provide treatment. As such, facilitating timely user interactions with the patient record can improve the usability of the EMR [36]. However, without understanding how users interact with an EMR, and the specific points of interaction, it is difficult to determine the possible set of opportunities to enhance its usability.

Almost a decade ago, Chen and Cimino [11,12] began to address this issue by using log file analysis to study the clinical information systems from a user-centered perspective. This work analyzed the way that EMR users access the patient's record to understand the

users' interaction with the elements that influence clinical decision making, such as characteristics about the users and patients [11]. An association role-mining framework was applied to learn the interactions that are associated with each other. For instance, in 10% of the user sessions (between login and logout), they found that the user opened pages on the main laboratory, the patient laboratory results, the radiology main, and the patient radiology reports. Moreover, they used a sequential pattern discovery technique to detect the order of clinical data types that the users viewed or invoked. For example, using data from New York Presbyterian Hospital logs, they discovered that the abdominal ultrasound result is viewed after a liver function test. It was also shown that there were specific ordered patterns associated with the utilization of clinical forms and the actions that transpire between the start and end of the clinical system user's interaction session. However, the interaction patterns do not describe the system from a patient-centric view at which the treatment process over multiple care providers revolves.

2.2 Frequent Pattern Mining

In this section, we describe what constitutes a frequent pattern and the main algorithms for pattern discovery.

Pattern mining discovers a set of common attributes shared among objects in a dataset [61]. The first usage of pattern mining was for analysis of market basket data to infer customers' purchasing habits and create a plan to increase the sales [20]. Since then, researchers have used frequent pattern mining to analyze interactions and association between events for wide range of environments. Consider several examples in market analysis, 70% of the people, who buy Jane Austen's Pride and Prejudice, buy Emma in the following month [61]. In EMR interactions, Cimino and colleagues [14] studied the way the users interact with the clinical systems to redesign the insertion and increase the usage of information buttons, which are applications that provides health information resources to clinical system users based on specific links.

Another successful implementation for frequent pattern mining is in similarity search of complex and structured datasets such as event logs, transaction sequences, and images [20]. In such datasets, the volume of the data is huge and searching for an object can take

a long time. Hence, summarizing the objects by their frequent patterns can reduce the dimensionality of the search space and decrease computational cost when searching [20].

For datasets that consist of sequences of categorical data, there are numerous algorithms to discover frequent patterns such as SPAM [3], PrefixScan [45], and SPADE [61]. In 2001, Zaki developed SPADE [61] that convert the dataset of sequence Identification Numbers (ID), the event time, and the events that happened at that time into a vertical database that consists of tuples in the form <itemset : (sequence_ID, event time)> [61]. Implementing the vertical data format reduces the number of database scans to discover frequent patterns [61].

2.3 Process Mining

Process mining corresponds to the automated discovery of ordered behavior in the form of models that characterize an organization and its performance [22,32]. Process mining is applicable to any event, or time, based systems such as business process management (BPM) systems (e.g., filenet), enterprise resource planning (ERP) systems (e.g., Microsoft Dynamics NAV, and SAP), and hospital information systems (e.g., Epic, ChipSoft) [22,32,30]. All of the aforementioned systems share one feature in common - the storage and analysis of events associated with a process in an electronic format along with the event time in logs. Each event represents a defined step that is performed by a specific actor or originator [22,30].

The goal of process mining is to infer a process model and project it against an intended, or an *a priori*, model [22,32,59]. Thus, process mining is positioned at the intersection of data-oriented techniques (i.e., data mining and machine learning) and workflow management to answer "What if" questions and verify the current steps of a process [57]. The following subsections describe some of the techniques that have used in processing mining.

2.3.1 Process Mining Algorithms and Petri Net Representation

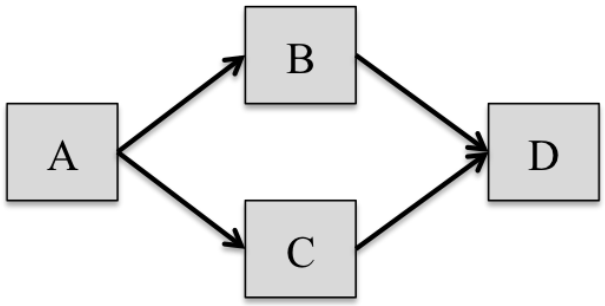
The most popular process mining algorithms are the α-algorithm, heuristic algorithm, and genetic algorithm. All of these algorithms provide the same type of model representation,

which is a petri net. The α-algorithm scans the event log to find the order relations between events [58]. The heuristic algorithm expresses the main behavior in the events log and attempts to reduce the effect of noise [60]. The genetic algorithm tries to mimic the process of evolution to tackle the duplicate entries and incomplete data problem [55].

A petri net is a mathematical and graphical modeling tool that describes the information system process [39]. It provides a graphical representation for the concurrent, sequential and asynchronies similar to the flow charts, which may represent the sequential patterns in the system. Figure 2.1 shows an example of an event log and its corresponding petri pet representation. Here each box represents a state or a task and the direction of the arrow indicates the sequence of events. The splitting out from a state is similar to an ″OR″ relation, where any one of the next states may happen. For example, after state ″A″ event ″B″ or event ″C″ may happen but both of them will be followed by event ″D″. The petri net simulates the concurrent activities or events in the system; however, it may provide a graph that is difficult to interpret when the system is complex, and has a substantial number of loops.

Process ID	Event Value
1	A
1	B
1	D
2	A
2	C
2	D

(a) Table of processes and their events



(b) Petri net for processes flow in table (a)

Figure 2.1. A petri net example, where the processes in (a) a tabular form are represented as a (b) flow chart.

2.3.2 Process Mining and Sequence Alignment

Comparing two or more sequences of items (numbers or letters) has been used in numerous domains, including speech recognition and molecular biology [9]. Multiple sequence alignment (MSA) was designed to find the optimal alignment for a set of

sequences through a series of edit operations, such that all the sequences have the same length and share the highest number of items at the same position [4,17]. The alignment procedure typically starts by overlaying the most similar sequences and gradually adds the next most related sequences [52]. MSA can be used on any set of sequences provided the vocabulary of terms (e.g., DNA, RNA, or a natural language sentence) is well-specified. More recently, MSA has been adopted by other fields such as process mining, and outlier detection in a set of sequences [9,22].

MSA has mainly been used in the fields of computational biology and bioinformatics fields where the goal is aims to discover biological and evolutionary processes of different organisms [27]. Discovering patterns across the biological sequences is essential to understanding evolutionary processes [13]. In support of this goal, researchers have used MSA methods to discover the conserved components of sequential categorical systems.

2.4 Process Mining in Healthcare

A healthcare organization is, essentially, an event-based system where care providers manage patient transition from one point of care to another, treat patients to manage a disease, or communicate with other parties for payments, supplies, or other actions. Understanding the healthcare process and discovering current flows of care can assist in improving the process and detecting bottlenecks [6,32]. Over the past several years, there have been an increasing number of studies into the application of process mining in healthcare.

2.4.1 Process Mining using Petri Net

Most process mining studies in healthcare have relied upon event logs (such as those affiliated with admission, discharge, registration, and radiology or laboratory orders) and the people who were involved in the care steps to represent the process [23,30,48]. Process mining techniques that use the actions of people and administrative tasks provide a representation for the organizational process.

Mans and colleagues [30] performed process mining on gynecological oncology workflows in an academic medical center in the Netherlands. First, they used heuristic

mining to generate a petri net for gynecological patients using event logs. Figure 2.2 depicts their petri net after preprocessing the data. Notice the complex spaghetti-like structure, which is difficult to read and directly interpret.

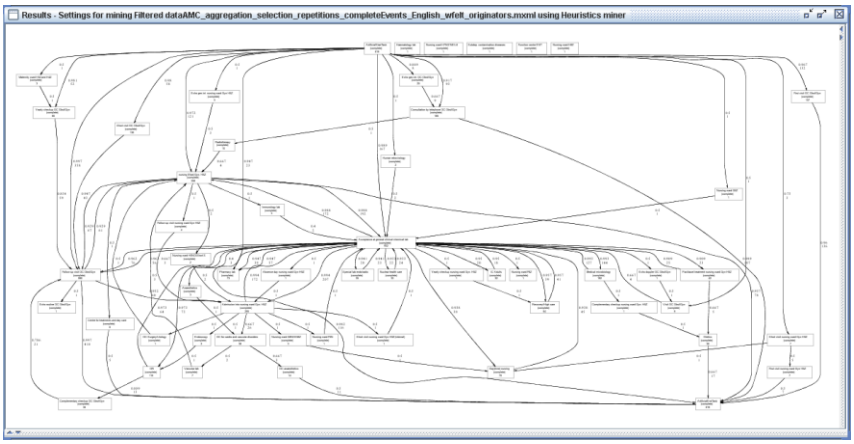


Figure 2.2. Petri net with a spaghetti-like model for a gynecological oncology workflow in a Dutch hospital reprinted from [30].

To obtain more informative models, they analyzed the clinical workflow from three perspectives:

1) *Control flows* in which they investigated the pathways that patients went through via visits. In this model, they only provided the type of visit and clinical visit pathways. The generated model lacks the details about social and clinical events and focused only on how the patient flows from one type of visit to another.
2) *Organizational and social networking* in which they investigated how collaborations and interactions between different departments at the hospital come together to treat patients with a specific diagnosis.
3) *Performance models* in which they aligned the events according to their relative time from the patient's admission to provide measures associated with admission duration, the number of events in each case, and patterns associated with certain treatment sequences. They represented the output in a dotted chart which represented the patient clinical path as a series of dots. Each dot represents an event that happened to the patients and different events had different colors. They tried to detect frequent patterns of clinical events from the chart using only visual inspection.

The three studies model form the clinical process, however, they did not discuss whether those models could be combined to obtain a complete clinical model. This would be difficult because each model represented the process in a different way.

Few investigations have used data from actual medical treatments, such medications [19,33,47]. Mans and colleagues [33] applied process mining on ischemic stroke treatments in four Italian hospitals (hospital names were not mentioned). In doing so, they attempted to identify the clinical pathways that the patients followed and contrasted the processes between the hospitals. They used the data from collected by stroke units in the four hospitals. They used process mining to produce petri net for the patients' treatment in stroke unit. They compared two hospitals' petri net and found a difference in the treatment protocol. Specifically one of the hospitals gave an antihypertensive medication while the other one did not. They concluded that one of the hospitals is a research facility since its petri net included neuroprotection which is a therapeutic protocol. This study was limited since it did not use data about the medications were administered, or were laboratory tests ordered by the Emergency department. These features, as our investigation shows, are critical to modeling ischemic stroke treatments.

2.4.2 MSA in Process Mining

Bose and van der Aalst [5,6,22] discuss the potential for MSA in discovering common patterns, processes, and deviations in the process. In [6], MSA is applied to mine processes in a variety of settings, such as telephone repair, rental allocation by a rental agency, and building permit requests in Dutch municipalities. In each process, they discovered common patterns or subsequences that the processes shared, as well as rare instances or deviations that some sequences exhibited. For example, in the telephone repair model, one of the sequences exhibited included a rare event that did not exist in other repair sequences. The deviated event was a failure to perform the first phone inspection because the customer was not at home.

In the medical domain, Bouarfa and Dankelman [7] introduced an MSA algorithm to align and recognize patterns in the sequence of instrument usage in laparoscopic cholecystectomy surgeries. They demonstrated that surgical process sequences could be

generated from event logs and applied MSA to discover the surgical workflow and outlying sequences. For example, one of the sequences included extra clipping and cutting of the cystic artery. They found that all the surgical sequences included an outlier, which is an item that the surgical sequence has in a position but the common alignment does not have it at that position. They justified their findings by claiming that the surgical workflow varies from patient to patient.

2.4 Challenges and Limitations in Clinical Workflow Mining

Lang and colleagues [24] evaluated different process mining approaches by detecting the process pathway in a radiology department. They created metrics and formulated equations to assess the output by measuring the completeness of the generated model, the ability to deal with noisy data, the ability to distinguish different processes, and the ability to deal with fuzzy entries and end points. They found that most of the process mining algorithms failed to either: i) discover the process models or ii) the discovered process models did not match the real known pathways [24]. Different factors contribute to the challenges in applying process mining to healthcare systems.

1- **Highly Dynamic.** Medical knowledge changes daily and can have a dynamic and complex nature [30,32], while the main actors in the healthcare system rotate. Different clinical teams or actors may treat different patients with the same disease with a slight variation. In addition, the patient's demographics, health insurance, and health status can increase the variability of treatments.
2- **High-Dimensionality and System Complexity.** Interactions in the healthcare system are non-linear and multivariate [26]. Different levels and types of people are involved in the treatment process, physicians including ranging in department and specialty, nurses, family of the patients, and the patients themselves. Moreover, a patient's demographic factors and health conditions (e.g., comorbidities) can induce a personalized set of treatment steps.
3- **Ad Hoc and Self-Organized.** The clinical systems can be described as emergent event-based self-organized systems, in which the occurrence of events shifts that care protocol [26]. Healthcare providers act according to their experience, knowledge, and current patient situation which varies from one case to another.

All of these factors can increase the variability in a patient's treatment and different order sets of care actions.

4- **Level of Abstractions.** The data stored in EMRs is highly diverse and may be generated by different subsystems in the healthcare organization. For instance, data may be derived from administrative records, clinical information systems, and medical devices [32]. While different data types can be combined to describe what happened to patients specifically and, more generally, in the organization, the data may exhibit different levels of abstraction. For instance, administrative data has a high-level of abstraction that considers the main checkpoints such as admission, discharge and registration. By contrast, clinical data has a different level of abstractions, some of the data have low abstraction (e.g. clinical notes), some data have average abstraction level (e.g. timestamp of ICD9 daily code includes the date only).

5- **Missing Data and Poor Documentation Procedures.** Although there is growing attention to improve clinical documentation, there remains a mismatch between current documentation interfaces and the application flows that support them [29]. Automating the completion of fields, providing copy-paste capability, and switching from an unstructured to a structured format can facilitate the clinical documentation process [49]. However, the mismatch between the documentation applications, the type of the care flow based on urgency, and the required accuracy, reusability and readability contributes to poor documentation [49]. When a certain document or data that describes a certain treatment step is missing or ill-presented, it could lead to improper modeling of the patient treatment. For instance, when applying a process mining strategy, it could suggest that patients with similar ailments and treatments had a different care plan, when, in reality, they all went through the same steps.

CHAPTER 3

DISCOVERY OF FREQUENT PATTERNS AND FORMATION OF CLINICAL PATHWAYS

In this chapter, we present a methodology to discover the common behavior in treating patient in specific cohort. First, we formalize the problem of treatment process discovery. Second, we introduce Treatment Mining using Frequent Sequential Patterns (TM-FSP), which is a multi-step process mining model to learn from EMRs to form the clinical pathways associated with the treatment of a specific diagnosis. It consists of three phases, as Figure 3.1 depicts:

Clinical Sequence Formation (Phase 1): First, it uses medications and laboratory tests to form the clinical sequence for each patient in the cohort.

Filtering and Dimensionality Reduction (Phase 2): Second, it discovers the frequent patterns using frequent pattern mining, and uses frequency as a proxy to reduces the dimension of sequential medical data by filtering the sequences into generated sequence frequent patterns.

Sequence Alignment and Clustering (Phase 3): Third, it clusters the filtered sequences based on their similarity. Each cluster is subject to MSA, and finally, TM-FSP represents the common actions among patient subtype (i.e. cluster) as a workflow.

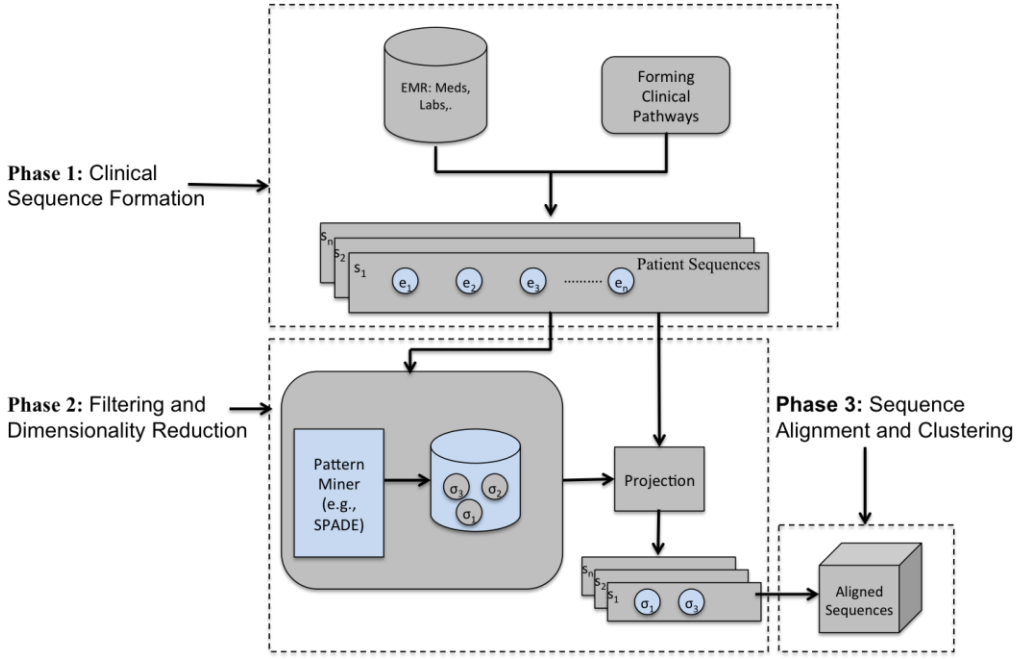


Figure 3.1. TM-FSP approach to generating clinical pathways for a cohort of patients. Phase 1: Formation of clinical sequences *s* using clinical events $\{e_1, e_2, .., e_n\}$. Phase 2: Reduction of the domain dimensionality using pattern set σ and projection of the sequences along the patterns. Phase 3: Generation of the clinical pathways via the application of MSA.

3.1 Treatment Process Discovery

The treatment for an admitted patient consists of performing clinical actions, such as the ordering of laboratory tests and medication administration. The healthcare providers know the main steps to treat the patients; however, the treatment process can be highly dynamic [28]. As a consequence, healthcare providers may not realize the different steps performed for a given set of patients nor the frequency of clinical actions among the patients in the cohort. We propose TM-FSP to discover the frequency and the order in which clinical events transpire.

The following sections describe the phases of TM-FSP. Each section starts with the main objective or high-level overview of the phase, then it provides a detailed description of the corresponding phase.

3.2 Phase 1: Clinical Sequence Formation

In this phase, TM-FSP takes a set of patients P and forms the clinical sequence for each patient. Let $C = \{c_1, c_2, \ldots, c_j\}$ be the set of single clinical events that were performed on the patient (e.g., admission, medication, laboratory order, and radiology scan). Clinical events may happen at the same time; hence, the set of clinical events $E = \{e_1, e_2, \ldots, e_i\}$ where e_i is the set of single clinical activities that were performed at t_i such that $e_i = \{c_1, c_2, c_i\}$.

We define the patient clinical sequence as the clinical events that healthcare providers performed to treat the patients from a specific disease. TM-FSP obtains the clinical events dataset E from a patient dataset P, and forms the clinical sequences. A clinical sequence s for patient $p \in P$ is the ordered clinical events such that $s_p = \{e_1, e_2, \ldots, e_n\}$, where e_i is a clinical events that happened at time t_i and $\forall i \neq j$: $t_i < t_j$ (*i.e.*, e_i happens before e_j). For each patient, we formed the clinical sequence s_p. The output of Phase 1 is a set of clinical sequences $S = \{s_1, s_2, s_3, \ldots, s_p\}$.

3.3 Phase 2: Removing Noise By Dimensionality Reduction

The core clinical pathway consists of common steps that were applied to treat the patients. The clinical environment is highly dynamic and contains a lot of variables, thus the data is vulnerable to noise and events that are not part of the clinical process (e.g., users can access a patient's record to check laboratory result status, medication or combinations made by the hospital for several patients). To dampen noise and reduce the dimensionality of the data, we invoke a frequent pattern mining method to uncover the number of dimensions necessary to represent the set of sequences.

3.3.1. Detecting Frequent Patterns

The first step in Phase 2 is discovering the events that happen frequently in the sequence set S and events that occur together in the same sequence across all the sequence set. Here, we formalize terms that will be invoked to describe methods in TM-FSP later.

Each sequence in S can be represented by events or subsequences such that $s' = (e_1, e_2, \ldots, e_i)$ that forms the sequence. Given clinical sequence S, the support of an event e_i or

subsequence of events is the number of sequences that include the event e_i or the subsequence s'. The minimum support threshold *minSupp* is the minimum number of sequences that include the event e_i or the subsequence s'.

A pattern σ is the subsequence of ordered events (e_1, e_2, ..., e_i) that occurs in at least *minSupp* of sequences in *S*. A subpattern σ' is a pattern that is part of another longer pattern. We say σ_i is a subpattern of σ_j if, and only if, $\sigma_i \subset \sigma_j$ (i.e., all the events in σ_j includes all the events in σ_i as well as other events). Similarly, we call σ_j a superpattern of σ_i. A pattern is considered to be closed if its support is greater than the minimum support threshold and none of its immediate superpattern has the same support. TM-FSP uses SPADE to generate the frequent patterns.

3.3.2. SPADE Mechanism

In this section, we describe how SPADE works. The input to SPADE is a dataset that includes: sequence id (SID), time of event(s) which is also considered as the event id (EID), and the set of events that happened at that time. To generate frequent patterns, SPADE creates a vertical id-list dataset. For each event, SPADE associates each sequence id that included the event with the time at which the event occurred in the corresponding sequence. SPADE starts by obtaining the most frequent events from the vertical id-list, where the support values for those events or generated patterns are more than the minimum support threshold. To create patterns that consist of two frequent events, SPADE joins the tables of two events and obtains the support value for the generated patterns. The events in the generated patterns do not have to be adjacent, other events can occur in between, but the events in patterns should occur in the same order as it occur in the sequences. Hence SPADE uses the timestamp with each sequence id to meet the order condition. To create patterns with three events or more, the same approach that SPADE uses to generate the patterns of two events using the joined tables of previous frequent patterns and frequent events tables.

For example, Figure 3.2(a) depicts an example of a dataset for which SPADE will generate the frequent patterns. In our example, we use 0.5 as the support threshold. SPADE generates patterns *A*, *C*, and *D* because their support values are all 1. In addition,

the vertical id-list that SPADE generates and are going be used to get the next patterns are Table *A*, Table *C*, and Table *D* as shown in Figure 3.2(b). To generate the frequent patterns of two events, SPADE joins the vertical id-list in Figure 3.2(b) taking into consideration the EID. To generate *A* → *C*, SPADE joins vertical list-id A and C, where EID (*A*) should be less that EID (*C*). Although A and C events were separated by B in sequence 1, but SPADE counted the pattern that counted that sequence in the pattern generation because SPADE only check the time condition which was met in sequence 1. When SPADE joins *C* → *D*, SPADE uses the events in sequence 1 and 2 to generate C → *D*, but SPADE does not use *C* and *D* events in sequence 3 because EID (*C*) is larger that EID (*D*). Hence, the support value of *C* → *D* is 0.67. Figure 3.2(c) depicts the frequent patterns generated in the second iteration. SPADE keeps generating patterns until it cannot generate any larger frequent patterns.

SID	EID	Events
1	10	A
1	20	B
1	30	C
1	40	D
2	10	A
2	20	C
2	40	D
3	20	A
3	30	D
3	40	C

(a) Dataset

A

SID	EID
1	10
2	10
3	20

C

SID	EID
1	30
2	20
3	40

D

SID	EID
1	40
2	40
3	30

(b) Vertical ID List for 1 event patterns

A → C

SID	EID (A)	EID (C)
1	10	30
2	10	20
3	20	40

A → D

SID	EID (A)	EID (D)
1	10	40
2	10	40
3	20	30

C → D

SID	EID (C)	EID (D)
1	30	40
2	20	40

(c) Vertical ID List for 2 events patterns

Figure 3.2. Generating frequent patterns via the SPADE algorithm.

3.3.3. Obtaining Frequent Patterns Using SPADE

The first step of Phase 2 takes the clinical sequence set *S* and discovers the closed frequent patterns that occur in all clinical sequences. The output of this step includes the events that all sequences in *S* share. Moreover, the frequent events and patterns form the dimensions that can represent the sequences without losing the essential data that characterizes the cohort of the patients.

3.3.4. Sequences Projection

In this step, TM-FSP reduces the dimension of the sequences by representing them through the discovered closed patterns. For each sequence *s*, it goes through all of the patterns, starting from the longest and proceeding to the shortest (TM-FSP goes from the longest patterns to the shortest or from the superpatterns to subpatterns). If the sequence contains pattern σ_i, TM-FSP passes the events that form the pattern to the projected sequence. If there are two patterns that overlap, the events which the patterns fail to share will be a subpattern and caught later in the reduction process. In the final output, the projected sequence includes only the events of the frequent patterns. Hence, the output is a compressed form of the original sequences. Figure 3.3 depicts the sequence projection algorithm.

Algorithm Sequences Projection

Require: Clinical sequences set *S*, patient clinical sequence s_p, events in sequence *e,* patterns set σ

Return: Projected sequences S′

```
1:  sort patterns in σ by support value ascending and by number of events in pattern descending
2:  initialize S′ ← {}
3:  for each clinical sequence s_p in S set
4:      Initialize s_p′ to empty sequence
5:      for σ_i in sorted σ
6:          if events subset {e_1,e_2,...,e_j} in σ_i exist in s_p
7:              Add events subset of σ_i to s_p′
8:      end for
9:      add s_p′ to S′
10: end for
11: return S′
```

Figure 3.3. Pseudocode for the clinical sequence projection algorithm.

3.4 Phase 3: Process Mining Through MSA

In this phase, TM-FSP clusters the sequences, and applies the MSA on each cluster.

2.3.1 Pairwise Sequence Alignment:

The pairwise alignment is a special case of MSA where the number of aligned sequence is tow. The alignment process enforces both sequences to have the same length by inserting gaps to represent an insertion or deletion process (or indel process) that happened in the sequences. Sequences can be aligned in different ways. To provide the best alignment, an alignment score is generated. Needleman and Wunsch [41] proposed a dynamic programming to find the optimal alignment based on a score. The alignment score is calculated by matching score (number of matched element multiplied by the number of matched elements) and the mismatches scores (the number of mismatch multiplied by the corresponding mismatch score). The mismatch between two sequences is represented by a gap and corresponds to an insertion process in one of the sequences accompanied with a deletion process in the other sequence. In other words, the alignment score is:

$$score = \sum_{i=1}^{L} e_i$$

Where:

L is the length of the tow sequences after the alignment

e_i is the score at position *i*:

$$e_i = \begin{cases} S(s_{1i}, s_{2j}) if\ s_{1i} = a\ and\ s_{2j} = b \\ I(s_{1i}, s_{2j})\ if \begin{cases} if\ s_{1i} = a\ , s_{1j-1} = b\ and\ s_{1j} = - \\ if\ s_{1i} = a\ , s_{1j-1} = -\ and\ s_{1j} = b \end{cases} \end{cases}$$

such that:

$S(s_{1i},s_{2j})$ is substitution value. If $a = b$, then the substitution is 1. Otherwise, the substitution will be -1, and

$I(s_{1i},s_{2j})$ is indel value, which is the penalty for inserting a gap.

2.3.2 Progressive Alignment Approach

For a given set of clinical sequences $S = \{s_1, s_2, \ldots, s_i, \ldots, s_n\}$, the progressive alignment constructs a succession of pairwise alignment. The alignment happens between two sequences, one sequence and generated alignment, and between generated alignments. To guide the multiple alignment process, a guide tree is produced based on the edit distance between aligned sequences and the features used to characterize the sequences such as N-Gram which a similarity measure that counts the co-occurrence of the set of character in strings to categorize electronic texts [10]. The guide tree will place the most similar aligned sequences in the same branch [22,52]. To generate the guide tree, TM-FSP uses an agglomerative hierarchical clustering (AHC). The AHC is a bottom up approach that iterates through all the sequences and clusters the most similar groups (based on the defined similarity metric) together until all the sequences are merged into one cluster [51]. The AHC builds the tree by iterating through the sequences, and align the most similar pair of sequences together. Then, the AHC aligns the next sequence or the alignment of a group of sequence with the most similar alignment of group of sequences to generate a new alignment. Following the branch in the guide tree, the progressive alignment will align the sequences in the same branch then align the most similar branches together to generate a dendrogram.

3.4.3 Clustering Similar Sequences:

To cluster the alignments, TM-FSP picks the best cut-point in hierarchical clustering, which is the point that determines to which cluster each sequence belongs. TM-FSP uses different number of cluster starting from 2 clusters. For a given clusters number, the TM-FSP choses the cut-point in the dendrogram and assign each sequence to a cluster. TM-FSP iterates through number of possible clusters. TM-FSP picks the clusters number that generate the least sum of pairs which are the edit distances between each pair in the cluster, in all clusters.

3.4.4 Applying MSA on Generated Clusters:

For the generated clusters $C = \{c_1, c_2, \ldots, c_m\}$, where m is the number of generated clusters, TM-FSP apply the MSA on each cluster. Given a set of clinical sequences from the same cluster, the MSA process creates a set of aligned traces $\bar{S} = \{\bar{s_1}, \bar{s_2}, \ldots, \bar{s_n}\}$.

Each sequence has the same length L after inserting the gaps in each sequence such that: $|\overline{s_1}| = |\overline{s_2}| , \ldots , |\overline{s_i}| = L$, where $\overline{s_i}$ is the sequence s_i after gap insertion. The output of the MSA is the consensus sequence that represents the backbone of all aligned sequences.

CHAPTER 4

WORKFLOW MINING FOR ISCHEMIC STROKE TREATMENT

We wanted to test the TM-FSP performance and whether we can detect the common clinical pathway for specific cohort and discover the clinical pathways for subpopulation. This chapter reports on an experimental analysis of TM-FSP using ischemic stroke cohort and data derived from the electronic medical record system of Vanderbilt University medical center (VUMC). The chapter begins with a presentation of the patient cohort. It then describes the data we used to form clinical sequences. Next, it describes how we removed the noise and reduced the dimensionality of the data using frequent pattern mining. Afterwards, it discusses how we used MSA to find common behavior. Finally it reports the findings regarding the clusters of clinical sequences we got and the difference between them.

4.1 Ischemic Stroke Cohort

To evaluate TM-FSP, we selected a population of patients diagnosed with ischemic stroke. This phenotype was chosen because its treatment requires a well-defined and ordered treatment steps [1,64]. Specifically, these steps can be formalized as a guideline, which should be followed by clinicians to maximize the outcome for the patient. The American Heart Association publishes an updated stoke management guideline every couple of years [1]. The Brain Attack Coalition is a group of professionals who aim to minimize the disabilities and deaths associated with stroke [64]. Via their website, they provide a well formatted stroke management steps that summarizes the stroke treatment guidelines.

In a Dutch case study [33], process mining was applied to discover the treatment process for ischemic stroke patients while they were treated by the Neurology department. They obtained the petri net and compared between two Italian hospitals. However, the study did not discuss the common pathway of patients' treatment nor the how similar the clinical pathways in those two hospitals. Moreover, they applied the process mining on part or segment of the treatment which is the Neurology segments. In our experiment, we

used all the clinical events during the whole treatment to obtain the common clinical pathway.

The process by which patients were ruled into the cohort is shown in Figure 4.1. First, we restricted our analysis to a four-month period in 2010, such that all patients were admitted and discharged (or died) during this time period. Initially there were 67,300 patients admitted. Second, we excluded patients who failed to be billed for ischemic stroke, using the 434.91 ICD9 code. This yielded 145 patients. Third, we removed all patients who were admitted for a different reason, such as sepsis or skin melanoma and had a stroke during the hospitalization. These patients were excluded because the treatment protocol that was followed at the beginning does not necessarily match a stroke protocol. This process yielded a final data set of 133 patients.

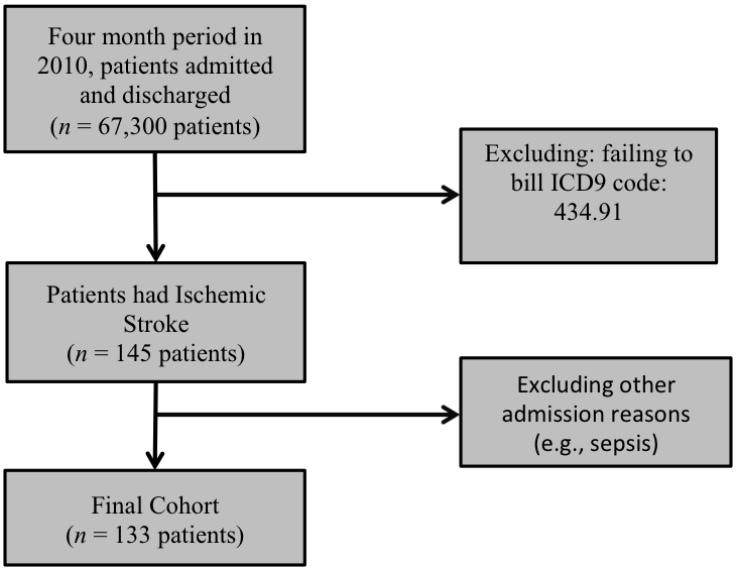


Figure 4.1. Cohort formation for ischemic stroke patients.

During their admission, the patients received different medical services based on the department that patient was admitted to or treated by. From the admission table, we obtained the services provided to the patients and the time windows during which they received the service. We defined the start and the end time for each service and obtained the events that transpired. Figure 4.2 depicts as example of segmenting the clinical sequence by the services that were provided to the patient. Table 4.1 provides a summary of the cohort and corresponding dataset. The patients were admitted, on average, for 241

hours. For 133 patients, 14 different hospital services were provided, where each patient might receive only one service or more during the admission. Table 4.2 provides the different services and the number of patients who received the corresponding service. We focused our analysis on two services: i) Emergency service (EMER) and ii) Neurology service (NEU). This was done for two reasons. First, both services are core to the provision of care for ischemic stroke patients. Second, the number of patients who received other services in other services is very small (15 patients and less). Of the 133 patients, 95 received service in the Emergency department and 88 received Neurology care. We refer to the period during at which the patients received care in a service as its stage.

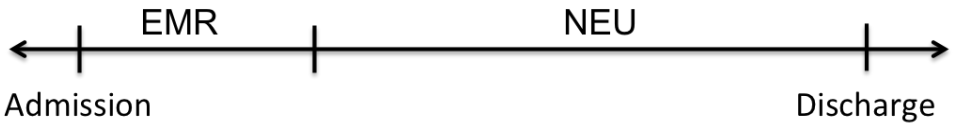


Figure 4.2. Example of segmenting patient's hospitalization based on provided service.

Number of patients	133
Average admission duration (hours)	241
Standard Deviation of duration (hours)	3.9
Minimum, Maximum of admission duration (hours)	2.0, 677.1
Patients with EMR service	95
Patients with NEU service	88

Table 4.1. Summary statistics for the ischemic stroke cohort.

Service name	Number of patients
Emergency	102
Neurology	88
General internal medicine	15
Pulmonary	5
Cardiac	5
Geriatrics	4
Trauma	2
Hematology	2
Vascular surgery	1
Obstetric general	1
Nephrology	1
Infectious disease	1
Emergency general surgery	1
Cardiac/thoracic surgery	1

Table 4.2. Frequency of services number provided for 133 ischemic stroke patients.

4.2 Clinical Event Sequences

Each patient was represented as a sequence of events documented in the EMR. Specifically, we represented each patient as a series of time-stamped medication and laboratory orders. The type of laboratory test was documented in the medical record. The medication table includes the National Drug Code (NDC), however, medications with the same effect have different NDC values. For example, METOPROLOL has a NDC value 51079025520, and CARVEDILOL has a NDC value 51079093020. Both drugs have different names and different NDC values but both are Beta Blockers. Hence, each medication was assigned its class according to the National Drug File – Reference Terminology (NDF-RT) [62]. In the VUMC laboratory table, each row represents a single item that is a part of a laboratory type. Each laboratory item has a battery value which consists of three sections: i) laboratory type abbreviation ii) more specific information about the laboratory type iii) a number that identify the laboratory performed for a specific patient. Table 4.3 provides an example of laboratory items and battery field value. Multiple items may share the same battery which means that all the items belong to the same laboratory type for the same patient. To obtain the laboratory type to which an item belongs, we take the first section of the battery field value. In the provided example, the lab type for all the items is basic metabolic panel (BMP). Table 4.4. provides a summary of the total number and variety of medications and lab tests ordered.

The dataset has 882 medications that group into 182 classes, and 630 laboratory tests items that were grouped into 241 laboratory types.

Lab item	**Battery Field**
AN-GAP	**BMP** BasMet XXXXXXXXXXXX
BUN	**BMP** BasMet XXXXXXXXXXXX
CO2	**BMP** BasMet XXXXXXXXXXXX
Ca	**BMP** BasMet XXXXXXXXXXXX
Cl	**BMP** BasMet XXXXXXXXXXXX

Table 4.3. An example of the laboratory items and battery values used to obtain the laboratory type.

	Medication	**Laboratory**
All patients	16220	57477
Distinct number	882	630
Class or type number	182	241

Table 4.4. Summary statistics regarding medication and laboratory classes.

4.3 Frequent Patterns

As noted in Chapter 1, a workflow is composed as the common items or patterns that are conserved across. We applied the SPADE algorithm [61] to obtain the frequent association patterns with a support of at least 40%. Applying SPADE generated 2,020 closed patterns. Table 4.5 shows some closed patterns that have the highest support values. One of the most important steps for stroke patient is checking the PT before administering the Anticoagulant. However, PT → Anticoagulant has support 0.6 and it is less than Anticoagulant support value 0.75, which might indicate that not all the documented Anticoagulant was preceded by documenting ordered PT test. From 182 medication classes, 7 were found to be frequent: i) Anticoagulants, ii) Antilipemic agents, iii) Beta Blocker/Related, iv) Insulin, v) Non-Opioid Analgesics, vi) Opioid Analgesics, and iii) Potassium. From 241 laboratory types, 12 were found to be frequent: i) A1C, ii) Blood Panel (BP), iii) CPD[**C**omplete Blood Count (CBC) / **P**latelet Count/ **D**ifferential], iv) Creatinine, v) Glucose, vi) Lipid, vii) Creatine Kinase, viii) the Partial thromboplastin time (PTT), ix) Prothrombin time (PT), x) Urine test, xi) INR or PT test, and xii) Troponin.

Pattern	Support
<{BP}>	0.977
<{PT}>	0.864
<{BP}→ {BP}>	0.848
<{GLB}>	0.826
<{MB}>	0.811
<{TRI}>	0.811
<{PT}→ {BP}>	0.788
<{GLB}→ {BP}>	0.773
<{ANTICOAGULANTS}>	0.750
<{BP,MB}>	0.742
<{TRI}→ {BP}>	0.742
<{BP}→ {ANTICOAGULANTS}>	0.742
<{CRE}>	0.720
<{NON-OPIOIDANALGESICS}>	0.720
<{BP}→ {NON-OPIOIDANALGESICS}>	0.705
<{MB}→ {BP}>	0.705
<{BP}→ {GLB}>	0.697
<{A1C}>	0.689
<{MB,TRI}>	0.689
<{BP}→ {BP}→ {BP}>	0.682
<{ANTICOAGULANTS}→ {ANTICOAGULANTS}>	0.682
<{BP}→ {ANTICOAGULANTS}→ {ANTICOAGULANTS}>	0.674
<{UA}>	0.667
<{PT,TRI}>	0.659
<{LIP}>	0.652
<{GLB}→ {BP}→ {BP}>	0.652
<{BP}→ {BP}→ {ANTICOAGULANTS}>	0.652
<{GLB}→ {GLB}>	0.636
<{PT}→ {BP}→ {BP}>	0.636
<{BP,MB}→ {BP}>	0.636
<{BP,PT}>	0.629
<{BP}→ {BP}→ {NON-OPIOIDANALGESICS}>	0.621
<{CRE}→ {BP}>	0.621
<{GLB}→ {TRI}>	0.614
<{BP}→ {ANTICOAGULANTS}→ {BP}>	0.614
<{PT}→ {NON-OPIOIDANALGESICS}>	0.606
<{MB,TRI}→ {BP}>	0.606
<{PT}→ {ANTICOAGULANTS}>	0.606

Table 4.5. Most Frequent patterns generated by SPADE.

In our study, we picked 0.4 as the minimum threshold for SPADE to generate the frequent patterns. Picking a threshold higher than 0.4 generated no more than 511 patterns. We attempted to use a threshold lower than 0.4 but the SPADE R-package execution stopped and it might happen because the number of generated patterns was very large. Hence, for future direction, if SPADE cannot generate patterns using small threshold value (e.g., 0.2), we can use 0.4 because it is the value that explains higher percentage of data variability.

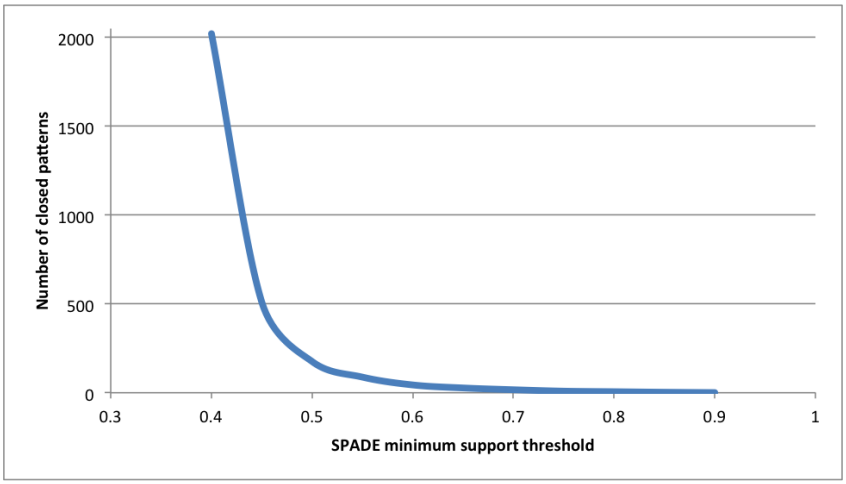


Figure 4.3. Relation between SAPDE minimum threshold and the number of generated closed patterns that explain data variability.

TM-FSP projected each patient's event sequence onto the set of frequent items to obtain a reduced (or denoised) view.

	Original	**Projected**
Minimum length	4 events	2 events
Maximum length	1,540 events	833 events
Average length	131.82 events	69.75 events
Standard Deviation	187.13 events	98 events
Total Number of events	18,151 events	10,885 events

Table 4.6. Sequences before and after projection for 133 ischemic stroke patients.

Table 4.7 summarizes the subsequences in the Emergency and Neurology services. The subsequences length in the Neurology service is, on average, by 62 events, with a

standard deviation of 102.97, which is much higher than the deviation of subsequences in Emergency services, which has a standard deviation of 11.55.

	Emergency subsequence	**Neurology subsequence**
Minimum length	1 event	2 events
Maximum length	83 events	833 events
Average length	5.66 events	62.11 events
Quartile of length (25%, 50%, 75%)	(2, 3, 4) events	(2, 30, 67) events

Table 4.7. Statistics for subsequences in Emergency and Neurology services.

4.4 Detecting Common Behavior using MSA

To apply MSA, TM-FSP uses the ProM6.3 tool, an open source framework for process mining [63]. We ran a series of investigations and experiments to sequences alignment. We performed the alignment on:

1- The original sequence without segmentation or applying any tokenizing criteria.
2- The projected sequence without segmentation or applying any tokenizing criteria.
3- The projected sequences for events that happened in the first 24 hours. This time period was selected because it is the most critical window in treating a stroke patient.
4- Projected sequences segmented into service subsequence: Clinical systems are event-based. The performed clinical events depend on the department and healthcare provider interacting with the patient. Hence, we created the subsequences based on the type of the service in which the event happened.

We aligned the original sequence (without removing the infrequent patterns or items) and the projected or denoised sequences. Then, we segmented the original and projected sequences based on the service time. The ProM6.3 tool provides a number of different options to i) select the sequence features to align the sequences, ii) pick the number of the clusters to group the sequences based on the best clustering score, and iii) align the sequences in the cluster with each other. For sequence features, we selected i) Individual events which align items that have exactly the same value, ii) *K*-Gram: In the alignment feature, we used $k = 4$.

Figure 4.4 depicts a cluster of aligned sequences for events that transpired on the Emergency service. Each row represents the treatment sequence for a patient after inserting the gaps to align the sequences on the common items and equalize the length of the sequences. Instead of using the item name in the aligned sequence, the PROM6.3 tool creates a mapping table that represents each distinct item in the sequences by a symbol. The symbol can be a lower case alphabetical letter, upper case alphabetical letter, or combination of alphabetical character and number depending on the number of required symbols to represent the items in the sequences. The number of mapping character depends on the number of distinct items in all sequences. The dash ″-″ represent an inserted gap. Table 4.8 shows the corresponding sequence items and the mapped character for the dataset used in part of study (in some parts of the experiment, we had to use more symbols to represent sequences items).

Sequence Item	**Alignment Character**
Insulin	A
Blood Panel	B
AIC	C
Troponin	D
Antilipemic Agents	E
Creatinine	F
Anticoagulants	G
CPD	H
Prothrombine	I
Glucose	J
Lipid	K
Creatine	L
Urine	M
Beta Blocker	N

Table 4.8. Character mapped to sequences' items in Figure 4.4.

At the bottom of Figure 4.4, the generated output incudes the common sequence that represents all the aligned sequences. In the left column, sequences (cases) of one of the Emergency service clusters, which are highlighted with gray, are either an empty sequence that did not include any event or sequences that lack no more than one of the events in the common sequence.

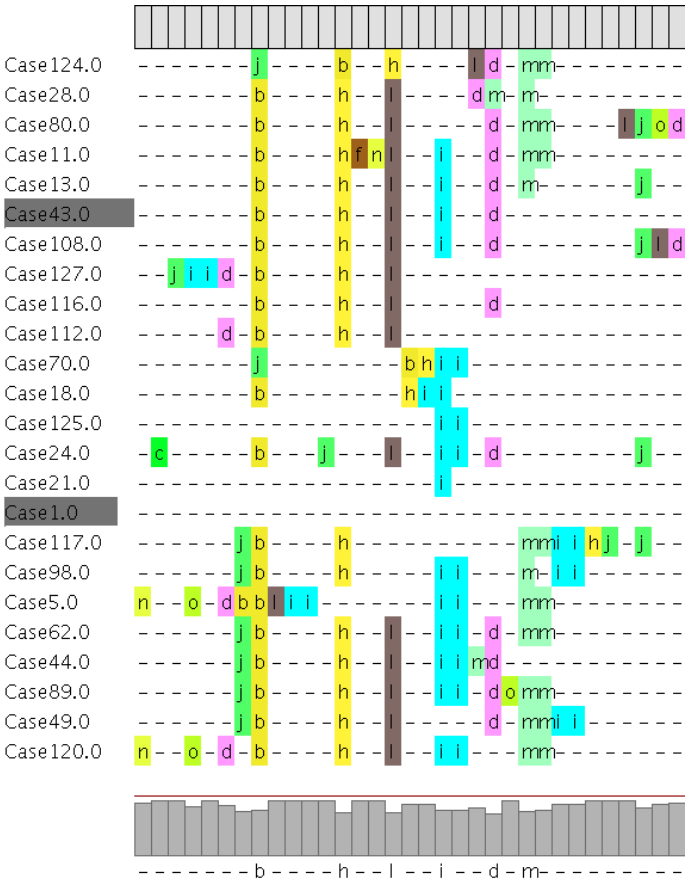


Figure 4.4. A multiple sequence alignment for one of the EMER service clusters.

4.4.1 Alignment Without Segmenting the Sequences

PROM was unable to align the sequences when we aligned the original sequence without removing infrequent items nor segmenting. This is because the dataset consists of 423 distinct events which is high number of events to be aligned especially if there are so many rare events. However, when we reduced the dimension or filtered the sequences without segmenting the sequence, and aligned them using MSA discovered two clusters. In Original_Cluster 1 the common sequence (i.e., output of MSA) contains 6 items, 3 of which are associated with a Blood Panel test, two with Insulin, and the last item is

Glucose. The MSA output for Original_Cluster 2 was empty, implying the items were too sparse.

The sequences for events that happened within the first 24 hours have, on average, 10 events with a standard deviation of 6.7. The smallest sequence included 2 events while the longest sequence included 29. We applied MSA on sequences that included events that happened for the patient within the first 24 hours. This yielded 4 clusters, each with a different common sequence. The alignments of clusters for sequences for the first 24 hours are is:

24Hours_Cluster 1: *Creatine Kinase* $\Rightarrow$ *Creatine Kinase* $\Rightarrow$ *Troponin* $\Rightarrow$ *Glucose* $\Rightarrow$ *Insulin* $\Rightarrow$ *Glucose*

24Hours_Cluster 2: *Creatine Kinase* $\Rightarrow$ *Glucose* $\Rightarrow$ *Insulin* $\Rightarrow$ *Glucose* $\Rightarrow$ *Insulin* $\Rightarrow$ *Glucose* $\Rightarrow$ *Insulin*

24Hours_Cluster 3: *Creatine Kinase* $\Rightarrow$ *A1C* $\Rightarrow$ *Non-opioid Analgesics*

24Hours_Cluster 4: *CPD* $\Rightarrow$ *Creatine Kinase* $\Rightarrow$ *INR* $\Rightarrow$ *INR*

24hours_Cluster 1 included Creatine Kinase, Troponin, Insulin, and Glucose, while 24hours_Cluster 2 included only Creatine Kinase, Glucose and Insulin. The common sequence for 24hours_Cluster 3 included only Creatine Kinase, Non-opioid Analgesics and A1C tests. Finally the common sequence for 24hours_Cluster 4 had CPD, Creatine Kinase, and INR. By comparing the alignment for all the clusters, the only element that all the clusters share and can be aligned at is Creatine, which one of main laboratory tests for ischemic stroke but not the most important laboratory nor the only one.

4.4.2 Alignment of Segmented Sequences Based on Service

In the Emergency service, TM-FSP discovered common behavior in three clusters. The main behavior that extracted was as follows:

$$Glucose \Rightarrow Blood\ Panel \Rightarrow Creatine \Rightarrow INR \Rightarrow Troponin$$

The common sequences for all Emergency clusters include laboratory tests implying that other diseases (e.g., hyperglycemia, hemorrhagic stroke) and check other patient's health conditions such as heart condition.

In the Neurology service, the sequences were longer, on average, by 56 events. MSA found three common sequences and the three clusters shared the following common sequence:

Blood Panel $\Rightarrow$ *Non-opioid Analgesics* $\Rightarrow$ *Anticoagulant* $\Rightarrow$ *Blood Panel* $\Rightarrow$ *Anticoagulant* $\Rightarrow$ *Non-opioid Analgesics*

4.5 Events Distribution Per Cluster

In both Emergency and Neurology services, the patient sequences were grouped into three clusters. Table 4.9 summarizes the number of patients in the generated clusters. In the Emergency service stage, the majority of the patients were grouped in EMER_Cluster 1, while the proportion of patients per cluster has different distribution of the Neurology service. It is worth mentioning that the patients who were in EMER_Cluster 1 during the Emergency stage could be in a different cluster during the Neurology service stage.

We obtained the distribution of clinical events in each cluster to recognize the difference between the clusters that belong to the same service. Figure 4.5 depicts applying the MSA on one of the Emergency clusters and the alignment of those sequences. To create Figures 4.5, we obtained the support value for each event in the corresponding cluster. For example, in Figure 4.5, we computed the support values for Blood Panel (BP) by dividing the number of patients in EMER_Cluster 1 for whom BP was assessed, on the number of patients in EMER_Cluster 1 (i.e. BP proportion of patients in cluster X= number of patients who had BP in cluster X/number of patients in cluster X). Figure 4.4 summarizes the distributions for the patients/events in the Neurology service in the three clusters. From the Emergency distribution in Figure 4.5, one of the clusters exhibited a set of clinical events that did not exist in the other two clusters. The EMER_Cluster 1 included all the clinical events and it is the only one that included Beta Blockers,

Anticoagulant, and Insulin, and all of those medications have low support value. Moreover, cluster one has the highest support value for Troponin test.

Service	Cluster 1	Cluster 2	Cluster 3
Emergency (95 patients)	80	3	12
Neurology (88 patients)	51	12	25

Table 4.9. The number of patients in each cluster for the Emergency and Neurology stages of care.

Figures 4.5 and 4.6 show the clinical events distribution in the Emergency and Neurology services respectively, and Table 4.10 includes a description for the acronyms that appears in the mentioned figures. As EMER_Cluster 1 and Firgure 4.5 depict, Troponin has support value of 0.9 which is one of the highest values among all clusters in Emergency service. One the other hand, the CPD support value in EMER_Cluster 1 has the lowest support value compared to EMER_Cluster 2 and EMER_Cluster 3.

The Neurology service yielded three clusters. As Figure 4.6 depicts, Insulin and Beta Blocker were not prescribed to the patients, and the lipid and glucose tests had the higher support in NEU_Cluster 2. In NEU_Cluster 1 the blood panel test, glucose test, and Insulin tests have the highest support, while NEU_Cluster 3, blood panel test, non-opioid analgesics, and A1C test have the highest support.

From the distributions of clinical events in the Neurology and Emergency services clusters, we want to know whether the proportion of patients for each event is statistically significant or not. For a specific event (e.g., BP or Anticoagulant), we used a Chi-square test and compared between the observed number of patients who had the event in the sequence versus the expected value. Table 4.11 lists the calculated Chi-Square values for the Emergency service. Statistical significance was only observed in Cluster 1 in which all of the values are statistically significant except Beta Blocker, Anticoagulant, and Insulin. All the computed Chi-square values were more than 5.9 threshold for degree of freedom equals 2. For the Neurology service, Table 4.12 includes the Chi-Square values for the patient proportion for a specific event. In EMER_Cluster 1, and Insulin was statistically significant. If we look at the Insulin support value shown in Figure 4.5, it can

be seen that most of the insulin administration occurred at EMER_Cluster 1. Insulin is statistically significant value in all the Neurology clusters. In NEU_Cluster 2, the Chi-Square for Beta Blocker is statistically significant. From Figure 4.6, it can be seen that Insulin and Beta Blocker were not given to the patients in NEU_Cluster 2, while the expected value for the patient proportion who should be given those medication should be higher than zero (as indicated from the Chi-Square value).

Description of Clinical Event	**Acronym in Figures**
Anticoagulants	AGU
AntiLipemic Agents	ALP
Beta Blockers	BB
Blood Panel	BP
Creatinine	CRE
Glucose Blood	GLB
Insulin	INS
LIPID	LIP
Creatine Kinase	MB
Triponin	TRI
Urine	UA
Coumadin	COU
Non-Opioid Analgesics	NOPD
Opioid Analgesics	OPD
Potassium	POT

Table 4.10. Abbreviations of events in Figures 4.5 and 4.6.

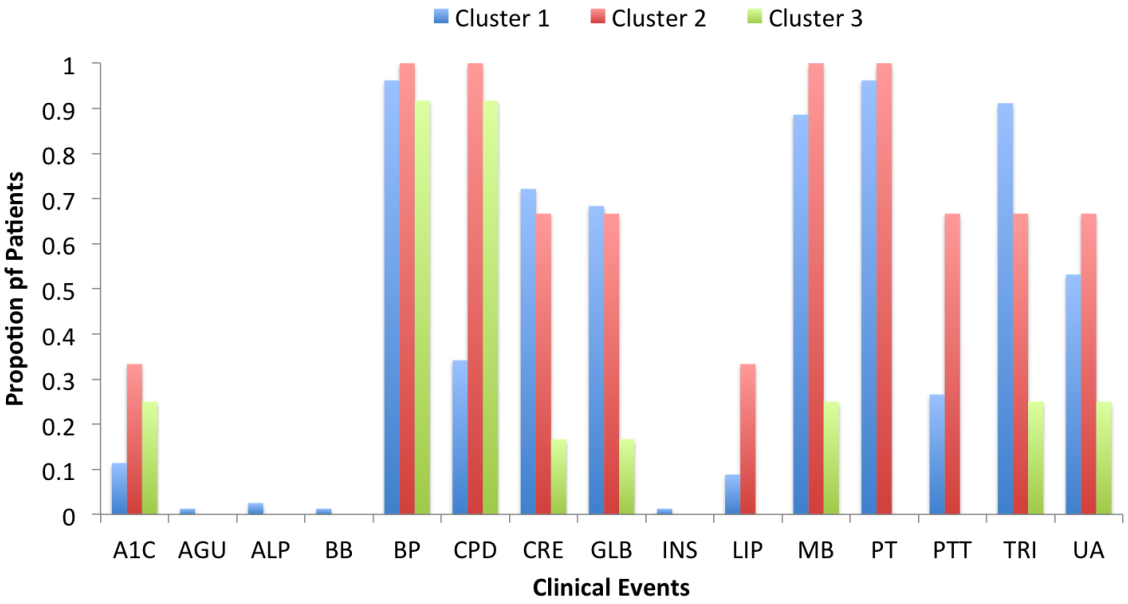


Figure4.5. Clinical event distribution for Emergency Service clusters.

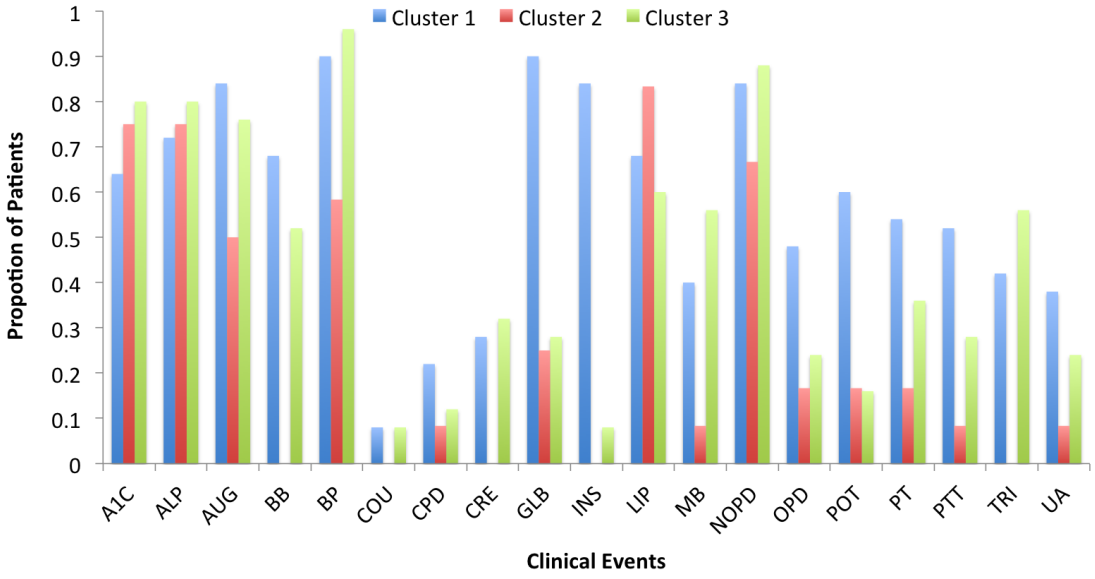


Figure 4.6. Clinical event distribution for Neurology Service clusters.

	Cluster 1 (χ^2, p-value)	Cluster 2 (χ^2, p-value)	Cluster 3 (χ^2, p-value)
A1C	0.30 (0.43)	37.03 (0)	0.250 (0.4)
AGU	0.03 (0.49)	3 (0.11)	0 (0.5)
ALP	0.07 (0.48)	**6 (0.03)**	0 (0.5)
BB	0.03 (0.49)	3 (0.11)	0 (0.5)
BP	0.02 (0.50)	**264.03 (0)**	0.92 (0.32)
CPD	1.48 (0.24)	**117.07 (0)**	0.92 (0.32)
CRE	0.78 (0.34)	**179.02 (0)**	0.17 (0.46)
GLB	0.69 (0.35)	**170.02 (0)**	0.17 (0.46)
INS	0.03 (0.49)	3 (0.11)	0 (0.5)
LIP	0.02 (0.50)	**22.04 (0)**	0 (0.5)
MB	0.73 (0.35)	**222.03 (0)**	0.25 (0.44)
PT	1.62 (0.22)	**231.04 (0)**	0 (0.5)
PTT	0.18 (0.46)	**65.06 (0)**	0 (0.5)
TRI	0.99 (0.31)	**227.02 (0)**	0.25 (0.44)
UA	0.22 (0.45)	**137.03 (0)**	0.25 (0.44)

Table 4.11. Chi-Square Values for Events in Emergency Clusters.

	Cluster 1 (χ^2, p-value)	Cluster 2 (χ^2, p-value)	Cluster 3 (χ^2, p-value)
A1C	0.27 (0.44)	0.04 (0.49)	0.35 (0.42)
ALP	0.05 (0.49)	0.00 (0.50)	0.09 (0.48)
AUG	0.31 (0.43)	1.14 (0.28)	0.003 (0.50)
BB	1.81 (0.20)	**6.48 (0.02)**	0.02 (0.50)
BP	0.04 (0.49)	1.16 (0.28)	0.21 (0.50)
COU	0.09 (0.48)	0.83 (0.33)	0.04 (0.49)
CPD	0.66 (0.36)	0.55 (0.38)	0.40 (0.41)
CRE	0.15 (0.47)	3.03 (0.11)	0.45 (0.40)
GLB	**5.67 (0.03)**	2.77 (0.13)	4.91 (0.04)
INS	**11.05 (0.002)**	**6.07 (0.02)**	**8.96 (0.01)**
LIP	0.000 (0.50)	0.43 (0.40)	0.23 (0.45)
MB	0.001 (0.50)	3.04 (0.11)	1.55 (0.23)
NOPD	0.01 (0.5)	0.38 (0.41)	0.08 (0.48)
OPD	1.71 (0.21)	1.32 (0.26)	1.11 (0.29)
POT	4.19 (0.06)	1.77 (0.21)	3.89 (0.07)
PT	1.22 (0.27)	2.01 (0.18)	0.34 (0.42)
PTT	2.14 (0.17)	2.90 (0.12)	0.79 (0.34)
TRI	0.04 (0.49)	4.83 (0.05)	1.55 (0.23)
UA	1.10 (0.29)	1.87 (0.20)	0.29 (0.43)

Table 4.12. Chi-Square Values for Events in Neurology Clusters.

4.6 Validation

We validated why we used Phase 2 of TM-FSP to remove infrequent patterns. We focused on validating the model using sequences segemented by service. To do so, we applied MSA on the original sequences (the sequence without applying SPADE) in the Neurology and Emergency services. In one of the Neurology service clusters, the alignments contained only blood panel test, CPD, and antilipimec agents, while the MSA could not align the sequences in the second cluster. In emergency clusters, the sequences did not align at any sequence in one of the clusters. However, the alignment produced a common sequence that included Blood Panel test, Glucose, and CPD.

To compare these alignments to those based on TM-FSP, we calculated the edit (Levenshtein) [40] distance between sequences within the same cluster. We calculated the edit distance. For the Emergency service that included the frequent patterns, the edit similarity values for EMER_Cluster 1, EMER_Cluster 2, and EMER_Cluster 3 are 0.599, 0.382, and 0.511, respectively. For the projected sequences in the Neurology service, the edit similarity values for the four clusters were: 0.72, 0.71, 0.64. The edit similarity values in Emergency service clusters are less than the edit similarity values in Neurology service clusters. Looking at Table 4.4, it can be seen that the sequences in the Neurology service were, on average, longer than the sequences in Emergency service; moreover the difference in the mean between sequences in Neurology and Emergency service is statistically significant with p-value less that 0.001 and confidence interval (-83.39, -34.12). Using long sequences will give us clusters with higher dissimilarity values and more general common behavior.

CHAPTER 5

DISCUSSION

The results suggests that applying process learning on the whole treatment process will provide only a general overview and will skip certain important interactions or cases that affect the treatment process. Dividing the treatment into stages, based on the type of event that were applied, can provide us more insight about the process and the main factors that may alter the treatment. Moreover, extracting the processes from similar patients generates a patient-centered flow that considers the standard process as well as the required variation in the treatment.

The frequent items generated by phase to TM-FSP matches the main laboratory test and the main medications that are recommended by the American Heart Association for stroke patients [1]. Using SPADE gave us the opportunity to provide an unsupervised noise removing or providing the main key factors in the treatment process. The anticoagulants neither the Partial Thromboplastin Time (PTT) nor Prothrombin Time (PT) were not administered for everyone. Moreover, not every anticoagulant administration was preceded by PTT test. This might have happened because some of the patients did not receive the full course of treatment (e.g., cancelled admission, or death), inability to capture all the instances or, simply, a poor documentation process.

When we aligned for the original sequence (i.e., non-reduced sequence), we only obtained two clusters, one of which did not have any common aligned sequence. When we aligned the sequence of events that happened within the first 24 hours, the common sequences for the generated clusters did not have common events except for Creatine, which indicates that the treatment plan or clinical process during the first 24 hours is not the same among the clusters. However, when we segmented the sequences based on the type of the provided service, MSA generated the common alignment for all of the clusters. The clinical process during the Emergency service mainly consisted of laboratory test. During the Neurology service, all the common sequences for the clusters included anticoagulant, blood panels and on-opioid Analgesics. Hence, those clusters

follow, in general, the same path in the treatment, with a small variation or differences in each cluster regarding the type of events and the proportion of patients whose sequences included that event. Moreover, The Insulin and Glucose test were one of the main alignment points in some clusters, which matches the long closed pattern of Insulin and Glucose test that we got from SPADE. Hence, dividing the clinical plan based on events can provide more robust results than a division based on time.

One of the major findings in our work is the discovery of a subpopulation that did not take Insulin and a Beta Blocker. This is surprising because most studies mention a positive effect of Beta Blockers on ischemic Stroke patients [16]. However, there are reasons to avoid Beta Blocker because it can have negative metabolic effects, such as dyslipidemia and reduced glucose control [18]. In addition, in the cluster that did not have Beta Blocker, Coumadin and Insulin events (NEU_Cluster 2), we found that the lipid lab test has the highest support value compared to support values in NEU_Cluster 1 and NEU_Cluster 3. This finding might indicate that the patients in NEU_Cluster 2 exhibited lipid metabolism problem that caused the withholding of a Beta Blocker prescription.

Another finding in our work is the relation between the type of clinical event and the provided service for the patients. The stage at which events happen influences the next logical treatment step. During the Emergency service, the subsequences mainly consisted of lab tests, while during the Neurology service, the alignment included different events. In particular, the alignment of most Neurology clusters occurred at the anticoagulant, blood panel test, glucose and insulin. In some cluster, the sequences aligned at beta-blockers, A1C and lipid tests.

Another interesting find is the ICD9 distribution among Emergency clusters and it might be related to the clinical events distribution among Emergency clusters in Figure 4.4. For example, in EMER_Cluster 1, the ICD-9 code that has the highest support value is V71, which is *Observation and evaluation for suspected condition*. When the exact value for the patients in EMER_Cluster 1 was checked, it was found to be V71.7 which is *Observation and evaluation for suspected cardiovascular disease*. Moreover,

EMER_Cluster 1 was the only cluster that included medication orders and had the highest Troponin support value. The distribution of the codes, as well as the set of unique codes, for each cluster provides insight the differences in the distribution.

The difference in the events distribution, and specifically in prescribing medication, indicates that the order medications and laboratory tests performed by healthcare providers depends on the patient. For instance, the sequence in which the clinical events happen caused the TM-FSP model to group similar patients which is indicated by the event distribution and the statistically significant differences in Beta Blocker and Insulin prescription.

Limitations

This work is limited in several aspects. First, the size of the dataset was small. In some clusters, we did not obtain a sufficient number of patients to explain the variability in certain clusters. Due to the size limitation, our analysis of clusters significant using Chi-square was not the optimal option. In the contingency table that shows the clusters and the corresponding number of patients who had specific events (taking medication X, performing lab Y), the count in some cells was less than 5, hence the Chi-square may provide inaccurate results. For such cases, using Fisher exact test will be more appropriate. However, Fisher formula can be applied only on 2x2 contingency table and we might have more than two clusters. To find the clusters with significant value using Fisher test, we apply the test on each clusters pair where we form 2-way table (i.e., 2x2 table) from each two clusters values. Even if there is a Fisher test for larger table, it does not specify which of the clusters has significant value. To detect the cells that have significant value using more general approach, we can apply general log-linear regression on the table. The cell with significant coefficient has a value that is different from expected one. We could not apply the log-linear regression since the number of observations is insufficient to provide a significant analysis or acceptable results. Second, the clinical process is much more than medications and laboratory tests, it will be more informative to include various activities such as radiology tests, billing, clinical visit, admission, scheduling and the most important factor the healthcare provider themselves. The model yielded the treatment patterns and the alignment of those patterns, but we will

need different criteria to segment the clinical process-based on the clinical event boundaries, such as service duration. For example, the length of stay and the number of events is much higher in the Neurology service compared to the Emergency service. Hence the alignment in Neurology might miss important factors or events. Comparing long sequences will provide a general model and will be less sensitive to treatment variation in the subpopulations.

Third, we used MSA to obtain the common behavior in treating similar patients. However, MSA is highly sensitive to the order. Hence, we plan to investigate methods that will decrease the effect of order. For instance in some cases, the documentation for some clinical events might be postponed for an hour like the time at which the patient took a CT scan. Hence, the medications performing would be entered into the system before the CT scan was documented. However, the order of the events based on the time might indicate that the patient took the anticoagulant before taking the CT scan.

Forth, we want to compare TM-FSP with existing process mining approaches. In [33], Mans and colleagues applied process mining on ischemic stroke treatment process that was performed in Neurology department in a Dutch hospital and represented the process using petri net. Since we evaluated our model using similar cohort, it will be helpful to implement Dutch study approach that they followed and compare between the output of their approach and the result generated by our model. This comparison indicates whether our model provides more interpretable results compared to existing approaches. Moreover, representing the treatment process in our dataset using petri net allow us to compare between petri net and MSA representations and identify the cases at which each approach would be more helpful and appropriate. It is a way to validate the performance of our process mining model against existing one.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This thesis introduced an automated model to mine clinical process using the clinical events that the patient experienced. Using this model, we demonstrated that the technique provides a way to discover the current clinical process in comparison to other methods or other studies. Moreover, our investigation shows that the sequence of clinical events can identify subpopulations within the same cohort based on the order of clinical events that were performed on the patients. This can enable new ways to look at clinical processes, such as a mix of standardization and personalization.

We plan to continue developing TM-FSP via several directions. First, we plan to validate our model on larger dataset. Second, we plan to test TM-FSP on another phenotype, such as Myocardial Infarction cohort to ensure that our model is general and can be applied on other phenotypes. Third, we plan to include additional clinical events types and scale the approach. As a possible extension, we can include ICD-9 codes to describe the difference in the clusters or cluster the sequences based on clinical similarity.

REFERENCES

[1] Harold, Adams P., et al. "Guidelines for the Early Management of Adults With Ischemic Stroke A Guideline From the American Heart Association/American Stroke Association Stroke Council, Clinical Cardiology Council, Cardiovascular Radiology and Intervention Council, and the Atherosclerotic Peripheral Vascular Disease and Quality of Care Outcomes in Research Interdisciplinary Working Groups: The American Academy of Neurology affirms the value of this guideline as an educational tool for neurologists." *Circulation* 115(20) (2007): 478-534.

[2] Appleby, John, et al. "Variations in health care: the good, the bad and the inexplicable." *King's Fund Report London* (2011).

[3] Ayres, Jay, et al. "Sequential Pattern mining using a bitmap representation." *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002): 429–35.

[4] Bacon, David J., Wayne Anderson F. "Multiple sequence alignment." *Journal of molecular biology* 191(2) (1986): 153-161.

[5] Bose, Jagadeesh Chandra RP., and Van der Aalst Wil MP. "When Process Mining Meets Bioinformatics." *IS Olympics: Information Systems in a Diverse World*. Springer Berlin Heidelberg, (2012): 202-217.

[6] Bose, Jagadeesh Chandra RP., and Van der Aalst Wil MP. "Trace alignment in process mining: opportunities for process diagnostics." *Business Process Management*. *Springer Berlin Heidelberg* (2010): 227-242.

[7] Bouarfa, Loubna, and Jenny Dankelman. "Workflow mining and outlier detection from clinical activity logs." *Journal of Biomedical Informatics* 45(6) (2012): 1185-1190.

[8] Bowens, Felicia M., Patricia Frye A., and Warren Jones A. "Health information technology: integration of clinical workflow into meaningful use of electronic health records." *Perspectives in Health Information Management/Ahima, American Health Information Management Association* 7 (2010).

[9] Carrillo, Humberto, and Lipman David. "The multiple sequence alignment problem in biology." *SIAM Journal on Applied Mathematics* 48(5) (1988): 1073-1082.

[10] Cavnar, William B., and Trenkle John M. "N-gram-based text categorization." *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval (Las Vegas, NV)* (1994): 161-175.

[11] Chen, Elizabeth S., and Cimino James J. "Automated discovery of patient-specific clinician information needs using clinical information system log files." *American Medical Informatics Association* (2003): 145–149.

[12] Chen, Elizabeth S., and Cimino James J. "Patterns of usage for a web-based clinical information system." *Stud Health Technol Inform* 107(Pt. 1) (2004): 18-22.

[13] Chenna, Ramu, et al. "Multiple sequence alignment with the Clustal series of programs." *Nucleic Acids Research* 31(13) (2003): 3497-3500.

[14] Cimino, James J., et al. "Redesign of the Columbia University infobutton manager." *American Medical Informatics Association* (2007): 135-139.

[15] Cooper, Jeffrey D., Copenhaver James D., and Copenhaver Carolyn J. "Workflow in the Primary Care Physician's Office: A Study of Five Practices." *Information Technology for the Practicing Physician. Springer New York* (2001): 23-34.

[16] Dziedzic, Tomasz, et al. "Beta-blockers reduce the risk of early death in ischemic stroke." *Journal of the neurological sciences* 252(1) (2007): 53-56.

[17] Edgar Robert C., Batzoglou Serafim. "Multiple sequence alignment." *Current Opinion in Structural Biology* 16(3) (2006): 368-373.

[18] Fonseca, Vivian A. "Effects of β-blockers on glucose and lipid metabolism." *Current Medical Research & Opinion* 26(3) (2010): 615-629.

[19] Gotz, David, Wang Fei, and Perer Adam. "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data." *Journal of biomedical informatics* 48 (2014): 148-159.

[20] Han, Jiawei, et al. "Frequent pattern mining: current status and future directions." *Data Mining and Knowledge Discovery* 15(1) (2007): 55-86.

[21] Huang, Zhengxing, Lu Xudong, and Duan Huilong. "On mining clinical pathway patterns from medical behaviors." *Artificial Intelligence in Medicine* 56(1) (2012): 35-50.

[22] Bose, Jagadeesh Chandra RP., and Wil MP van der Aalst. "Process diagnostics using trace alignment: opportunities, issues, and challenges." *Information Systems*

37(2) (2012): 117-141.

[23] Kim, Eunhye, et al. "Discovery of outpatient care process of a tertiary university hospital using process mining." *Healthcare informatics research* 19(1) (2013): 42-49.

[24] Lang, Martin, et al. "Process mining for clinical workflows: challenges and current limitations." *Studies in health technology and informatics* 136 (2007): 229-234.

[25] Lau, Francis, et al. "Impact of electronic medical record on physician practice in office settings: a systematic review." *BMC Medical Informatics and Decision Making* 12(1) (2012): 10.

[26] Lipsitz, Lewis A. "Understanding health care as a complex system: the foundation for unintended consequences." *JAMA* 308(3) (2012): 243-244.

[27] Luscombe, Nicholas M., Greenbaum Dov, and Gerstein Mark. "What is bioinformatics? A proposed definition and overview of the field." *Methods of Information in Medicine* 40(4) (2001): 346-358.

[28] Malin, Bradley, Nyemba Steve, and Paulett John. "Learning relational policies from electronic health record access logs." *Journal of biomedical informatics* 44(2) (2011): 333-342.

[29] Mamykina, Lena, et al. "Clinical documentation: composition or synthesis?." *Journal of the American Medical Informatics Association* 19(6) (2012): 1025-1031.

[30] Mans, R. S., et al. "Application of process mining in healthcare–a case study in a Dutch hospital." *Biomedical Engineering Systems and Technologies*. *Springer Berlin Heidelberg* (2009): 425-438.

[31] Mans, R. S., et al. "Process Mining in Healthcare." *International Conference on Health Informatics* (2008): 118-125

[32] Mans, Ronny S., et al. "Process mining in healthcare: Data challenges when answering frequently posed questions." *Process Support and Knowledge Representation in Health Care. Springer Berlin Heidelberg* (2013): 140-153.

[33] Mans, Ronny S., et al. "Process mining techniques: an application to stroke care." *Studies in Health Technology and Informatics* 136 (2008): 573.

[34] Mant, Jonathan, and Hicks Nicholas. "Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction." *BMJ: British Medical Journal* 311(7008) (1995): 793.

[35] Mant, Jonathan. "Process versus outcome indicators in the assessment of quality of health care." *International Journal for Quality in Health Care* 13(6) (2001): 475-480.

[36] Middleton, Blackford, et al. "Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA." *Journal of the American Medical Informatics Association* 20(1) (2013): 2-8.

[37] Mount, David W. "Sequence and genome analysis." *Bioinformatics: Cold Spring Harbour Laboratory Press: Cold Spring Harbour* 2 (2004).

[38] Mulyar, Nataliya, Van der Aalst Wil MP, and Peleg Mor. "A pattern-based analysis of clinical computer-interpretable guideline modeling languages." *Journal of the American Medical Informatics Association* 14(6) (2007): 781-787.

[39] Murata, Tadao. "Petri nets: Properties, analysis and applications." *Proceedings of the IEEE* 77(4) (1989): 541-580.

[40] Navarro, Gonzalo. "A guided tour to approximate string matching." *ACM Computing Surveys* 33(1) (2001): 31-88.

[41] Needleman Saul B., Wunsch Christian D. "A general method applicable to the search for similarities in the amino acid sequences of two proteins." *Journal of Molecular Biology* 48 (1970) 443–453.

[42] Oreja-Guevara C, et al. "Clinical pathways for the care of multiple sclerosis patients." *Neurologia* (2010) 25:156–162.

[43] Panella, M., Marchisio S., and Di Stanislao F. "Reducing clinical variations with clinical pathways: do pathways work?" *International Journal for Quality in Health Care* 15(6) (2003): 509-521.

[44] Patel, Vimla L., et al. "Representing Clinical Guidelines in GLIF Individual and Collaborative Expertise." *Journal of the American Medical Informatics Association* 5(5) (1998): 467-483.

[45] Pei, Jian, et al. "PrefixSpan: mining sequential patterns by prefix-projected

growth." *Proceedings of the International Conference on Data Engineering* ICDE,01 (2001): 215–24.

[46] Peleg, Mor, et al. "The InterMed approach to sharable computer-interpretable guidelines: a review." *Journal of the American Medical Informatics Association* 11(1) (2004): 1-10.

[47] Perer Adam, Gotz David. "Data-Driven Exploration of Care Plans for Patients" *CHI'13 Extended Abstracts on Human Factors in Computing Systems* ACM (2013): 439-444.

[48] Rebuge, Álvaro, and Ferreira Diogo R. "Business process analysis in healthcare environments: A methodology based on process mining." *Information Systems* 37(2) (2012): 99-116.

[49] Rosenbloom, S. Trent, et al. "Data from clinical notes: a perspective on the tension between structure and flexible documentation." *Journal of the American Medical Informatics Association* 18 (2011): 181-186.

[50] Skinner, Jonathan. "Causes and consequences of regional variations in health care." *Handbook of health economics* 2 (2012): 45-93.

[51] Steinbach, Michael, Karypis George, and Kumar Vipin. "A comparison of document clustering techniques." *KDD workshop on text mining*. 400(1) (2000): 525-526.

[52] Thompson, Julie D., Higgins Desmond G., and Gibson Toby J. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic acids research* 22(22) (1994): 4673-4680.

[53] Unertl, Kim M., et al. "Describing and modeling workflow and information flow in chronic disease care." *Journal of the American Medical Informatics Association* 16(6) (2009): 826-836.

[54] Van der Aalst, Wil MP. "Challenges in business process mining." *Applied Stochastic Models in Business and Industry (Unpublished working paper)* (2010).

[55] Van der Aalst, Wil MP, de Medeiros AK Alves, and Weijters A. J. M. M. "Genetic process mining." *Applications and Theory of Petri Nets*. *Springer Berlin Heidelberg* (2005): 48-69.

[56] Van der Aalst, Wil MP, et al. "Workflow mining: A survey of issues and approaches." *Data & Knowledge Engineering* 47(2) (2003): 237-267.

[57] Van der Aalst, Wil MP. "Decomposing Petri nets for process mining: A generic approach." *Distributed and Parallel Databases* 31(4) (2013): 471-507.

[58] Van der Aalst, Wil MP, Weijters Ton, and Maruster Laura. "Workflow mining: Discovering process models from event logs." *IEEE Transactions on Knowledge and Data Engineering*, 16(9) (2004): 1128-1142.

[59] Van der Aalst Wil MP. "Process Mining: Discovery, Conformance and Enhancement of Business Processes." *Heidelberg Springer* (2011).

[60] Weijters, A. J. M. M., Van der Aalst Wil MP, and De Medeiros Alves AK. "Process mining with the heuristics miner-algorithm." *Technische Universiteit Eindhoven, Tech. Rep. WP* 166 (2006): 1-34.

[61] Zaki, Mohammed J. "SPADE: An efficient algorithm for mining frequent sequences." *Machine Learning* 42(1-2) (2001): 31-60.

[62] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/

[63] http://www.promtools.org/prom6/

[64] http://www.stroke-site.org/guidelines/tpa_guidelines.html "Accessed on: June 19, 2014"