Doctoral Thesis

A Data Driven Retrospective Study for Medication Strategy Analyses on Longitudinal Prescription Records

Purnomo Husnul Khotimah

August 2018

Department of Social Informatics
Graduate School of Informatics
Kyoto University

Doctoral Thesis
submitted to Department of Social Informatics,
Graduate School of Informatics,
Kyoto University
in partial fulfillment of the requirements for the degree of
DOCTOR of INFORMATICS

Thesis Committee: Masatoshi Yoshikawa, Professor
Tomohiro Kuroda, Professor
Kazuyuki Moriya, Professor

A Data Driven Retrospective Study for Medication Strategy Analyses on Longitudinal Prescription Records[*]

Purnomo Husnul Khotimah

Abstract

As the number of people with chronic diseases increases significantly, electronic medical records (EMR) has been accumulating big data of medication history. A retrospective study uses existing data that have been recorded. With the increasing attention on evidence based medical guideline, it is important to study the medication strategy from EMR to maintain the patient condition in chronic diseases, for example type 2 diabetes. The medication strategy may change over time because of a newly found adverse drug reaction or a newly released drug. Being able to understand these changes from a long-term medication history is essential in longitudinal studies, which are useful for learning the relationship between risk factors and the development of disease, and the outcomes of treatments over different lengths of time. Therefore, a clinical physician would benefit from these studies to develop an evidence based medical guideline. However, with the nature of big data, it will be difficult for clinical physician to conduct the study. Data driven tools can be used to conduct a data driven retrospective study to extract patterns of medication strategy out of the medication history.

One of prominent tools is frequent sequential pattern mining (FSPM). This method is proven useful in many area, including medical field to extract useful patterns. Our study focuses on analyses of medication strategy over a long-term medication history by observing the medication transition events. Although this study have been a long time interest for clinical physicians, previous clinical studies mainly concern on certain treatment strategies which is usually known knowledge using statistic methods. Using FSPM, previously unknown and useful information may be explored. On clinical field, existing FSPM studies focus

[*] Doctoral Thesis, Department of Social Informatics, Graduate School of Informatics, Kyoto University, KU-I-DT6960-26-9245, August 2018.

on short-term dataset. Therefore, those methods of short-term dataset analyses, may consider only medication's occurrences in the dataset. However, in long-term analyses, we may have to consider the medication duration because physicians need to wait for a certain period of time to see whether a medication is effective or not. Hence, previous methods result may contain spurious patterns resulted from short medication episode that have less meaningful for a clinical physician. We adapt the notion of time error margin ϵ for constructing the medication episode from prescriptions dataset which is provided by Kyoto University Hospital. Using 14 days as the ϵ value, we are able to compress the search space into 23.83% compared to the raw dataset.

Moreover, existing methods in medication history analyses using FSPM based their algorithm on Apriori, which features non-consecutive sequence and subset of itemset for candidate generation. These features may lead to a huge number of frequent patterns that inhibit clinical physicians to explore the result set. Furthermore, the patterns may have less meaning because clinical physician may not understand clearly about the medication strategy, such as which medication is stopped, or switched. Our mining methods named singleton mining and its extensions enable clinical physicians to get a finer grain result set compared the conventional method; for example, clinical physicians are able to understand with the way of diminishing of Sulfonylurea (SU) usage in the subpopulation with the monotheraphy of SU. Based on our result, after 2010, the replacement medication for SU in the subpopulation of patient previously having monotheraphy SU is DPP4-inhibitor (DPP4-i), whereas, in patients previously having any medication combination with SU, the replacement medication is not exclusively DPP4-i. Biguanide is also preferred by physician as medication replacement of SU in this subpopulation.

Furthermore, clinical physicians need a tool to explore the result set of medication strategy effectively. Conventional methods to present the result set is in tabular view based on certain ranking functions. In actual effort to present the result set, tabular view does not help much the clinical physician to get insight from the result. Using our mediation transition graph, clinical physicians are able to identify unfamiliar pattern compared with the recommendation from the medical guideline of type 2 diabetes treatment that is even though the patient condition is in ideal state the physician change the previous medication to medication combination using DPP4-i.

We have addressed several issues on medication strategy analyses from a long-term chronic medication history as the following identification of longer duration of medication episodes, observing medication transition events, generating a fine grain pattern that covers deeper clinical research questions and medication strategy visualization. Current proposed methods are applicable in a more general chronic condition where clinical physician needs to answer research question using a long-term medication history.

The following are several social-informatics impacts of enabling long-term medication history analyses by our method:

1. Improved health care. In practical side, our proposed method enables clinical physicians to analyze, long-term medication transition events, which hold important information such as the relationship between risk factors, health outcomes, or even adverse drug reaction. Therefore, the proposed method allows the advancement of medical knowledge and development of new treatments for chronic health problems.

2. Availability of new dataset. Our proposed method provide health care professionals with a new dataset in the form of medication episodes. This new dataset provide many usable information that could drive wider synthesis and analyses of data when connected with other dataset such as patient test result or medical billing.

3. Enhanced analyses to big data. Our proposes method enables more focus studies. A finer grained pattern enables analysts to investigate certain sub-populations of the entity in the dataset. This ability allows an individual profile to be compared with the sub population profile. Such kind of information is essential in personalized applications.

In summary, our proposed methods will not only impact clinical communities such as epidemiology societies, but also the entire field in health to benefit from comprehensive research studies.

Keywords: medication strategy, longitudinal analyses, chronic medication history, frequent sequential pattern mining

Contents

List of Symbols

α	A sequence
β	A sequence
δ	Minimum period threshold of a stable period
τ	Kendall's Tau coefficient
ϵ	Time error margin

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

Chronic diseases are health conditions that last long duration and generally slow progression [7]. Based on WHO documentation on noncommunicable diseases (NDC), the burden of chronic diseases is rapidly increasing worldwide [40]. In addition, the proportion of the burden is expected to increase to 57% by 2020. Furthermore, the chronic disease problem is far from being limited to the developed regions of the world. Contrary to widely held beliefs, developing countries are increasingly suffering from high levels of public health problems related to chronic diseases. Furthermore, $7 trillion is an estimated loss of productivity and price of health care without taking action over the next 20 years [41]. Therefore, chronic diseases pose a public challenge that undermines social and economic development of both developed and developing countries.

One serious, costly, and increasingly common chronic disease is diabetes. Diabetes is a condition characterized by chronic hyperglycemia due to deficiency of insulin action [51] The condition causes the increase of glucose in the blood. One of the prominent indicators is A1c level. The A1c level reflects the mean blood glucose level of during one to two months before the time when the blood sample was taken. The normal value of subjects with normal glucose tolerance is between 4.6% and 6.2%. People diagnosed with diabetes is indicated with the

A1c value of 6.5%.

Diabetes is also a progressive disease. It may lead to increasing risk factors for other conditions, such as heart disease, amputation, and kidney failure. Hence, diagnosing and treating the disease timely and appropriately reduces serious and costly complications and death. Currently, medical societies publish evidence based medical guidelines [51][50][38] in order to improve the healthcare quality and to give a comprehensive information about the disease, especially for physicians and patients. The effort, to find supporting clinical evidences, is requiring a clinician to conduct a retrospective study by investigating the previous treatment. Furthermore, because diabetes prevalence number is high and increased highly each year, Electronic Medical Records (EMR) has accumulates big data of type 2 diabetes patients medication history. Therefore, to analyze such a long-term and big data will be an exhaustive task. Hence, data driven longitudinal analyses play a key role in retrospective studies.

1.2 Motivation

Opportunities In Type 2 diabetes, the glucose metabolism disorder may worsen over time. Physicians are required to adjust the medication in order to manage the patients condition. Adjustments of medications in the prescription represent the physicians strategy and the patients pathway. In addition, the medication may cause adverse drug reactions (i.e., is an unwanted or harmful reaction experienced following the administration of a medication or combination of medication under normal conditions of use and is suspected to be related to the drug). For example, long term of Sulfonylurea (SU) medication may cause hypoglycemia in aged patients [34]. In this condition, physicians would modify the medication to minimize the reaction. Other case that may cause physicians to change the medication is the new released drug [26]. With the increasing attention on evidenced based treatment, understanding these adjustments over long term period will provide insights about medication strategies toward chronic conditions.

Pharmacotherapy in Diabetes Type 2 Diabetes medical guidelines provide information for general practitioners and people with diabetes about the diseases, patient condition management, and the standard treatment care, including the pharmacotherapy. However, the way medical guideline present the recommendation may differ from one and the other. For example, the guideline of Japan does not give a particular recommendation with which pharmacotherapy should start a certain treatment [51]. By contrast, guideline of America gives recommendation that the treatment should start with metformin [50]. Due to such conditions, longitudinal analysis of medication history can be used to capture the typical medication pattern of medical societies that use guideline with a more general recommendation and detect the outliers on the one that use guideline with a more comprehensive recommendation.

Moreover, one of diabetes treatment characteristic is that a physician needs to wait three months to find whether the medication is effective or not [51]. Therefore, in the case of longitudinal analysis, short term medication (less than three months) may represent noises where physicians still adjust the medication that may be irrelevant for long term medication strategy. Hence, we need to be able to identify medications with longer duration that associated with patient able to attain the ideal condition. We later, in section 2.1, refer this period of physicians waiting period as the *3 months rule*. Furthermore, the medication history is available in the form of prescriptions. The prescription nature are sometimes available in short duration due to the regulation. There are also gap and overlaps between the prescription durations. Therefore, an appropriate strategy is necessary for the data driven analysis to be able to mine the proven effective medications.

Physicians Adherence Towards Medical Guidelines Despite the medical guideline, the physician may follow the guideline recommendation or give alternative medication or treatment based on the patient condition and the physician experience. One of prominent study to investigate the doctor adherence is done by [53]. [53] analyzed the national guidelines for the management of type 2 diabetes to identify clinical conditions that are not covered or those for which the guidelines do not provide recommendations using data driven tools. [53] use C5.0 decision-tree learning algorithm to analyze patient records corresponding to each clinical condition from a database of type 2 diabetic patients treated at a hospital. The results shows that there are clinical conditions that do not have any recommendation in the previous medical guideline. However, those clinical conditions are in later medical guideline. From the study, we are able to say that there are rooms for medical guidelines to be improved. It is because a new medicine can be released or a new strategy towards patient condition can be developed.

The shortcoming of previous methods in studying medication history Several studies have used Frequent Sequence Pattern mining (FSPM) to conduct investigations on diabetes patients medication history. However, these studies used short-term data, which only include information on when the patient is hospitalized [5][14] and used the medication history as they are that is no preprocessing was done towards the raw data (after cleaning) [22][55]. These methods may only consider the occurrences of the medication. However, our study emphasis on analyses using long-term dataset. One advantage in using long-term data is that we are able to learn the medication strategies to maintain the patient conditions over a long period of time, which is a prerequisite. Such studies have been a long time interest for clinical physicians. However, previous clinical studies mainly concern on certain treatment strategies which is usually known knowledge using statistic methods [17] [16]. Furthermore, in a long-term medication strategy analyses, a characteristic in the pharamacotheraphy recommendation, such as the *3 months rule* should be considered. Moreover, previous studies used only monotherapy data [55]. By contrast, the diabetes type 2 pharmacotherapy varies from monotherapy to multitherapy. Hence, the scope of previous study is limited only for certain subpopulation of the diabetes patients.

To this end, we attempt to develop a practical strategy for data driven retrospective study of medical strategy analyses using long-term medication history in the form of prescription records.

1.3 Research Problem and Main Contributions

As mentioned in previous section, that electronic medical records (EMR) has been accumulating a big data which harbors insight and knowledge for a retrospective study towards the development of evidence-based medical guidelines. With the increasing attention towards evidence based treatment, a clinical physician is required to conduct retrospective study towards past treatments. Longitudinal analyses use long term medication history. These analyses benefit the clinical physician by giving information particularly about evaluation of the relationship between risk factors and the development of disease, and the outcomes of treatments over different lengths of time [12]. As we have mentioned in section 1.2 that most of previous study used short term dataset, which usally in the form of hospitalization dataset [53][5][22][55].

We are interested in studying the physician strategy in managing the patient with type 2 diabetes in a long-term. To conduct this longitudinal analysis, it is possible to use outpatient dataset, which medication (treatments) is being collected and patient condition (the outcome) is being monitored over long period of time. However, as mentioned in section 1.2 that the nature of this study is not appropriate for direct use. Therefore, preprocessing task is needed. In addition, we need to consider also chronic characteristic treatment, for example in type 2 diabetes, it is recommended to be stepwise and medication combination selection should adjusted with the patient condition. Thus, the order of the treatment and medication combinations are essential information for the clinical physician. This condition should be considered in modeling the pattern which should be extracted. Moreover, visualization of the result set should also be considered in order to provide an effective analysis by the clinical physician towards the medication strategy. Hence, we divide the task of our study into three parts, that is, preprocessing, mining activity, and visualization.

First, we list the limitation of our study as follows:

- The dataset at hand is in the form of prescriptions records of outpatients with diabetes. The prescriptions can be medication in the form of monotherapy to multitherapy.

- Unlike methods to study medication exposure that focus their studies to find out the medication duration, our prescription dataset contain the duration information. Therefore, we do not conduct an exposure duration estimation study.

- The medications used are varied from oral medicine to injection medicine. We exclude the injection type of medicine which require natural language processing.

- The dataset is limited only from one provider healthcare and not an integration from more than one healthcare provider. Therefore, the results in this study only represent Kyoto University Hospital's treatment characteristics.

- The variable values used in the experiments are respected with the type 2 diabetes diseases characteristics and the clinical pratice of Kyoto University Hospital.

- Tasks covered in this thesis are divided into three areas that is preprocessing, mining activity, and visualization.

And the following, we list main contributions of our study in longitudinal analyses of chronic medication strategy.

- **A framework for constructing medication episode from medication history.** In chronic longitudinal analyses, short term medication may represent noises that may represent a period of time when a physician tries to adjust the medication towards the patient condition. Therefore, we need to be able to reconstruct the medication history which originally in the form of short and repetitive prescriptions into a longer form of medication episodes (ME) (i.e., a period of time where a physician do not change the medication). The reason is that so we could identify which of the medication transition events that are useful for the analyses. In the medication history reconstruction, the previous method focused on the construction of a treatment episode (TE) (i.e., a series of temporally contiguous health care services related to treatment of a given spell of illness or provided in response to a specific request by the patient or other relevant entity [47]), to investigate drug utilization in relation to various drug taking related outcomes such as estimation of prevalence/incidence, compliance, and persistence [18]. This focus is different with our task, where in constructing TEs, adjacent prescriptions with different medications may assemble a TE. While in constructing ME, only adjacent prescriptions with same medications may assemble an ME. In addition, the previous method only consider mainly the gap duration for assigning adjacent prescriptions belongs to one TE. For our proposed method, we consider also the overlap duration in MEs construction. Our framework is able to transform the prescription dataset into medication episode that enables the observation on medication transition events and the identification of long term medication over a more compact dataset, which is a significant feature in data driven analyses.

- **The notion of singleton/full itemset in frequent pattern model.** In the case of chronic clinical condition, medications are given continuously over the patient's life time to maintain the quality of life. In order to adjust the medications, physicians consider the previous prescribed medications as one of the consideration besides the patient condition. Therefore, transition patterns between adjacent medications contain interesting information about the medication strategy, such as, which medication is added, stopped, or switched. One of prominent data driven tools to extract patterns from big data is frequent sequence pattern mining (FSPM). The conventional method of FSPM originally to solve market basket case that is to find out which items are frequently bought together [1]. Therefore, the method features non-consecutive sequence and subset of itemset in generating the frequent sequence candidate. These features result the frequent patterns does not contain a clear information whether itemsets (medication combination) in the patterns are adjacent to each other or not and whether a medication is not used or whether a medication is stopped or switched when an item is not in the itemset. Several constraint based FSPM was proposed to give additional criteria for the candidate generation such as the distance between the itemsets and the absence of item(s) from the itemset [44][33]. However, even using constraints, the conventional method unable to address the requirement for the information of the actual medication combinations. Our proposed notion of singleton/full itemset tackle this requirement. Moreover with incorporating this notion and the conventional method, our method enables a finer grain of frequent pattern result compared to the original conventional method.

- **Visualization of medication strategy.** In order to do longitudinal analyses, clinical physician needs to be able to explore the result set of the medication strategy. A medication strategy consists of clinical condition(s) that represent the physicians reasoning and medication transition event(s) the represent the physician's actions. Previous method in visual analytics focus their visualization to show the temporal relations between interval based events, such as A overlapping B or A followed by B [28]. This is different with the required information by the clinical physician. We developed two directed graph based visualizations. The first visualization is to show medication transition events combined with the patient conditions prior the transitions. This visualization method allows clinical physician to explore the k-top result set and derive conclusion about the reasoning of the medication transition events. On the second visualization, we provide a medication progression graph from monotheraphy to multitheraphy. The current visualization enables physician to understand the *adding* medication strategy.

1.4 Thesis structure

The structure of this thesis will be organized with respect to common steps in data mining that is data preparation, mining activity, and result presentation. Therefore, the rest of this thesis is organized as follows:

- Chapter 2

  This chapter describes the pre-processing step that is the medication episode construction framework. We introduce the adaptation of time error margin (ϵ) to construct medication episode from prescription dataset and propose the notion of duration threshold (δ) to identify stable periods (i.e., a period of time where a physician do not change the medication).

- Chapter 3

  This chapter explains about the mining method and introduced our proposed method that is named singleton mining method. Methods-comparison study and confirmatory experiment are conducted to highlight the strong points of the proposed method.

- Chapter 4

  This chapter describes the visualization methods developed to illustrate the medication strategy as the result of the mining method.

- Chapter 5

  The final chapter lists conclusions reach so far from the study conducted so far and some final remarks about a future work plan.

CHAPTER 2

Medication Episode Construction

2.1 Motivation

Prescription registries not only show a patient's medical history but can also be used as the information sources for drug utilization and pharmacoepidemiology analyses [49]. Many of the studies using prescription registries require the construction of treatment episodes [18]. One important aspect in treatment episode construction is determining which prescriptions are considered to belong to the same episode [45]. The process of medical history reconstruction from prescriptions into other forms, such as treatment episode, needs to be considered carefully. It is because once the process is complete, the outcomes of the later activities will be based on the extracted data. However, previous study relating to the usage of prescriptions datasets discussed only briefly about the medical history reconstruction. Other studies discussed the treatment episode construction focus on the drug exposure estimation because their datasets do not include duration information. Hence, a standard framework for performing medical history reconstruction out of prescription datasets is still lacking.

For chronic diseases, clinicians often require to perform longitudinal analyses of medical histories over the years. For example, one of the common chronic diseases is type 2 diabetes. The treatment characteristic is recommended to be patient-centered, that is, respectful of and responsive to individual patient preferences, needs, and values [54]. In addition, the treatment spans over the years throughout the patient's life. Therefore, a physician is required to develop a strategy to provide the best outcome not only for the short term but also for the long term [51]. Hence, for chronic diseases, it is necessary to assess the physician's long-term strategy. Our research specifically focuses on observing the medication transition events, that is, when the medication treatment changes from one medication to the next. Medication transitions are important not only to show the physicians' actions toward the disease progress (the patient condition changes), but also to demonstrate the treatment development as new medicines or techniques released over the years [43].

A framework for medical history reconstruction is required for enabling observation of the medication transitions from hospital prescriptions. This is because there are several issues in the nature of the prescription dataset. First, the periods of two consecutive prescriptions are sometimes unconnected and overlap with each other because a patient may come earlier or later than the appointment. Second, prescriptions generally have short duration due to some regulations, hospitalization or simply because of physician behavior. Third, many prescriptions are a continuation from previous medication when the patient achieved the target control assigned by the physician. Thus, we have to be able to express fragmented prescriptions into an aligned medication episodes (i.e., a period of time when a doctor prescribes the same medication continuously). Therefore, we can observe the medication transitions precisely only after constructing the medication episode.

Our next concern in prescription reconstruction is the prescription relations. As we previously mentioned, prescriptions may be unconnected (have a gap between each other), overlap or meet each other (connected). These are the possibilities when the dataset only includes monotherapy. This condition is because, in monotherapy datasets, patients having more than one medications at a time is excluded. However in a multitherapy dataset, there can be more than one prescriptions at one time which can have different medication with the currently taken by the patient. As a result, the prescriptions can have more possible temporal relations. For example, when a prescription has not been finished, a patient may visit a physician and received a new short-duration of prescription that does not last as long as the previous on-going prescription. This event will lead to previous prescription containing the new prescription. Hence in a multitherapy dataset, there are more prescription relations that need to be addressed.

A previous study by [18] attempted to construct treatment episode of an antidepressant treatment. The study proposed two treatment episode construction methods using a prescription time gap parameter. The first method does not add overlap duration of the successive prescription at the end time of the treatment episode, whereas the second method adds the overlap duration if the successive prescriptions belong to the same Anatomical Therapeutic Chemical Code (ATC). A treatment episode constructed by both methods in [18] is a period of time consists of connected prescriptions, which are previously separated by small gaps. Thus, in one treatment episode, there can be no changes in the medication (i.e., one treatment episode has the same medication) or there can be changes in the medications(i.e., one treatment episode has more than one medication). This is different with our goal of reconstructing the medical history that is to construct medication episodes out of prescription dataset. In addition, [18] only considered monotherapy by excluding patients who are prescribed more than one medicine at one time. This situation is in contrast with the nature of chronic medication that includes multitherapy prescriptions. In our study, the proposed method takes multitherapy dataset into account by using possible temporal relations between consecutive events that had been defined by Allen in [3] to address more possible prescription relations available in the multitherapy datasets.In addition, the concept of time error margin (ϵ) is employed to provide flexibility in assigning the temporal relation. As we will show in a later section, this variable is important for the medication episode construction in chronic disease analyses.

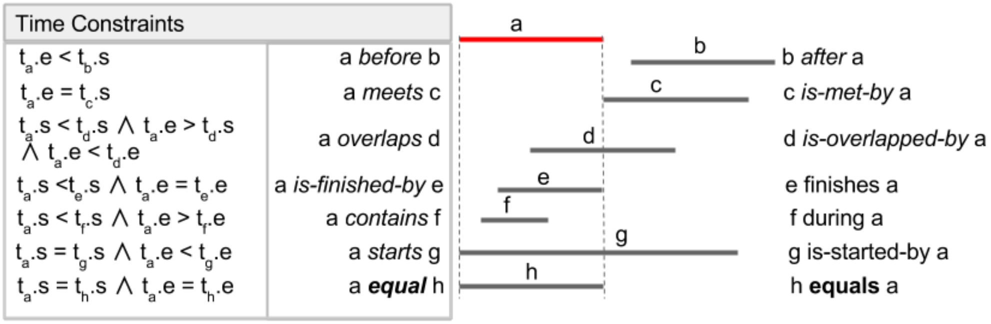
Furthermore, for the case of diabetes treatment, the physician often needs to wait three months to evaluate the effectiveness of the medication [51]. The recommended waiting time is 3 months which we refer to as the *3 months rule*. Based on the *3 months rule*, we hypothesis that medication episodes that have a duration at least 3 months and those with longer duration have more essential meaning for longitudinal analyses. This hypothesis is because longer duration medication indicates that the patient condition is reaching the control target assigned by the physician. Thus, we need to be able to observe the medication transition events between these type of medication episode, which later we refer to this medication episode as a stable period. Hence in addition, our proposed medication episode reconstruction is to enable the identification of such stable periods which has not been discussed in previous related studies.

This chapter will focus on the medication episode construction framework for chronic diseases which will be applied on a multitherapy hospital prescription dataset of T2DM patients. Specifically, we emphasize the usage of Allen's temporal relation and the time error margin in the framework, and the stable period generation.

2.2 Related Works

2.2.1 Time Interval Based Model Evolution

Our proposed framework in constructing medication episode is closely related with temporal relations between interval based events. The first one to address this topic is Allen in [3]. [3] introduced seven temporal relations (*before*, *meets*, *overlaps*, $is-finished-by$, *contains*, *starts*, and *equal*) with their inverses (13 relations in total) as shown in Figure 2.1. Temporal relations between interval based events enable the capturing or extraction of temporal knowledge out of natural language or relative information. For example, increasing medicine dosage *after* the rising of A1c value.



Time Constraints		
$t_a.e < t_b.s$	a *before* b	b *after* a
$t_a.e = t_c.s$	a *meets* c	c *is-met-by* a
$t_a.s < t_d.s \wedge t_a.e > t_d.s \wedge t_a.e < t_d.e$	a *overlaps* d	d *is-overlapped-by* a
$t_a.s < t_e.s \wedge t_a.e = t_e.e$	a *is-finished-by* e	e finishes a
$t_a.s < t_f.s \wedge t_a.e > t_f.e$	a *contains* f	f during a
$t_a.s = t_g.s \wedge t_a.e < t_g.e$	a *starts* g	g is-started-by a
$t_a.s = t_h.s \wedge t_a.e = t_h.e$	a ***equal*** h	h **equals** a

Figure 2.1: Allen's temporal relations with time constraints.

Allen's temporal relations has been used and developed widely. [25] added end time points (start and end time) of the interval to define the intervals constraint relations. Figure 2.1 displays the time constraints for each temporal relations. In [25], only seven temporal relations (the left part from Figure 2.1) are used so that one event is represented by only one temporal relations to avoid confusion.

[20] used temporal relation matrix to capture all possible relations in multivariate series datasets. The size of the matrix grew with the number of observed intervals. [42] introduced the notion of time error margin (ϵ) to define more flexible matchings between two interval based events. Hence, the temporal relations, which previously strictly constrained by start time and end time, have a more flexible relation by ignoring small differences in accordance with the ϵ value. For example, there are two instants of incident: event A has no gap from event B and event C has 1 day gap from event D. Hence, with using ϵ value of 5 days, both incidents have the relation of *meets*. This result is because the time constraint of the *meets* relation is relaxed by $\pm\epsilon$ (i.e., $t_1.e = t_2.s \pm \epsilon$).

This temporal relation model has been applied and developed for analyzing clinical data. However, the studies mainly focus on the temporal relation as the final objective of the study. For example, [32] used Karma Lego algorithm to find temporal interval relations pattern between the A1C and defined daily dosage (DDD) of diabetes medicine. In our case, we use the temporal relations as a tool in the medication episode construction to assign which rule should be applied for different prescription relations.

Furthermore, our framework also related with Moerchen's time series knowledge representation (TSKR) model, which was proposed in [31]. [31] showed that Allen's relations are not robust for noisy time series, ambiguous for one relationship may represent different conditions, and not easily comprehensible for one condition may be represented by different relationship [31]. TSKR was proposed to mine multivariate time series data by transforming the time series into interval symbolic series and finding the coincide intervals. Moerchen proposed the interval series as *tones* (i.e., observed parameters) and the interval coincide series as *chords*. In Figure 2.2, we are able to see three tones (A,B,and D). The three tones are resulting into four chords (A,AB,B,BD and B). Temporal relations of the chords will be represented as A → AB → B → BD → B. Such a representation is suitable, particularly for studies that observed more than two parameters because the number of all possible Allen's temporal relations will increase highly as the number of observed parameters multiplies. Hence, to observe medication transition events, *chords* model can be used to represent the overlapping prescriptions prescribed by the physicians as the medication episode.

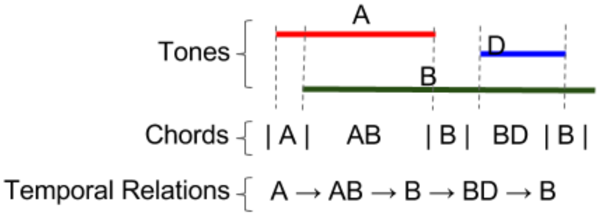


Figure 2.2: Moerchen's chords.

2.2.2 Prescription Reconstruction

In pharmacoepidemiology and drug analyses studies, patient's drug episodes are often assessed. The prescription registry as a data source often omits out duration. The available information generally includes the date of redeeming the prescription and the dispensed amount. Hence, approximating patients' actual drug use is needed. Several methods to estimate the duration of each prescription have been discussed in [19][37][36][18][45].

The first method consists of using the ratio of prevalence and incidence rate relationship and using this ratio as a constant period for each prescription. This is performed by assuming a constant use of dose (e.g., defined daily dose) or assuming other fixed amounts [19]. However, this method has been reported to have several caveats [19]. The second method is by using waiting time distribution [19][45]. The waiting time distribution (WTD) is a frequency distribution of first occurrence of drug use within a time window. In this second method, compensation for overlap and grace period are not considered because the duration is estimated from the maximum interval between prescriptions.The third method is by filling gaps between prescriptions[37][36][18]. By using the date of redemption and the dispensed amount, it is possible to define episodes of drug use. However, it is difficult to be sure if and when the dispensed are used.

Thus far, the aforementioned methods focused on the duration estimation rather than the prescription reconstruction. However, as mentioned in the previous section, our main idea in connecting prescriptions is the most similar with [18], which concatenated prescriptions to construct treatment episodes (i.e., prescriptions that are dispensed within the allowed gap that elapses after the expected end date of a prior prescription). Two methods of treatment episode construction based on the maximal gap were introduced, as shown in Figure 2.3. The first method on Figure 2.3.(a) does not add the overlap duration in the end of the expected end time of the treatment episode, whereas the second method adds the overlap duration when medication belongs to the same ATC group because the patient may come earlier as in Figure 2.3.(b). Hence, the second method affects the original gap to become shorter. Both methods introduced in [18] were applied for monotherapy dataset, which as the consequence prescriptions considered were prescriptions that next to each other either have a gaps or not, and overlaps to each other. In addition, the [18] study focused on comparing the effect of maximal gap variation on both methods.

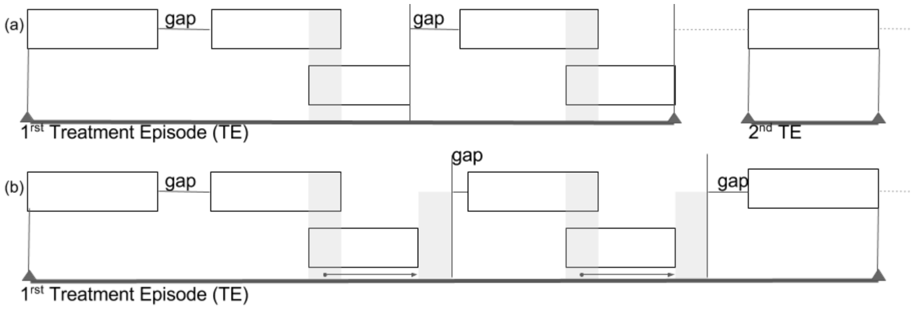


Figure 2.3: Two methods introduced in [18].

This is different with our study case, we are using a multitherapy dataset. Hence, there are more possible temporal relations between consecutive prescriptions compare to monotherapy dataset. Moreover, we consider the fact that a patient may come earlier or later than the appointment scheduled, which means successive prescriptions with short gap or short overlap should be connected as if the prescription were connected. By using this assumption, we use the concept of time error margin ($\pm\epsilon$) for not only to identify short gaps from longer gaps, but also to identify short overlaps from longer overlaps. Compared to two methods introduced by [18], our method seems to similar with the second method [18](the one that adds the overlap duration). In the second method [18], the successive prescriptions are considered overlapping regardless the overlap interval is (whether it is short or long overlaps). However, in our method, the prescription relations between two successive prescription are relaxed by $\pm\epsilon$. Hence, the successive prescriptions considered overlapping when the overlap duration is more than ϵ. Furthermore, our study focus is different from [18] that emphasize the study of comparing two construction methods (not adding overlap and adding overlap). Our study is focus on the ϵ variation in the generation of stable periods to enable observing the medication transition events. To summarize, Table 2.1

shows the property comparison between [18] and our study.

Table 2.1: Properties in our study compared with [18].

Property	Related previous study [18]	Our Study
Dataset	Monotherapy	Multitherapy
Parameter used	Maximal allowed gap	Time error margin (ϵ)
Allen's temporal relations	Fixed	Relaxed
Considered temporal relation	$before$, $meets$, $overlaps$	$before$, $meets$, $overlaps$, $is-finished-by$, $contains$, $start$, $equal$
Successive prescriptions turned into a *meets* relation	Prescriptions with a gap that is not more than predefined parameter value	Prescriptions with a gap or overlaps that is not more than predefined parameter value
Final result	Treatment episode construction	Stable period identification
Observed data behavior	median length of treatment episode, the number of patients' proportion based on their length of treatment episode	The generation of short and longer stable periods, the number of stable period sequence, and the number of medication transition events

2.3 Methodology

In this section, we describe the proposed framework of the medication episode construction for enabling the medication transition events between stable period. The description include the input (the prescriptions), the method (medication episode construction), and output (the stable period identification).

Diabetes medicine is classified into several types of medicines based on how they work. Table 2.2 shows medicine types with their medicine names.

Definition 2.1 *Medicine name* ***medName*** *is the proprietary name of the medicine. Each medicine name belongs to a single medicine type* ***medType*** *(i.e., medicine classification based on how the medicine works).*

Table 2.2: Medication types.

No	Medication Type	Medicine Names
1	Sulfonylurea (SU)	Rastinon, Euglocon, Daonil, Glimicron, Glimicron HA, Amaryl
2	Rapid-acting insulin secretagogues (RaIS)	Starsis, Fastic, Glufast, Surepost
3	α-Glucosidase inhibitors	Glucobay, Glucobay OD, Basen, Basen OD, seibule
4	Biguanides	Glycoran, Medet, Metgluco, Dibetos, Dibeton S, Melbin
5	Thiazolidinediones	Actos, Actos OD
6	DPP-4 inhibitors	Glactive, Januvia, Equa, Nesina, Tranzenta, Tenelia, Suiny
7	Combination	Glubes
8	Insulin	Novorapid, Apidora, Novolin, Innolet, Lantus, Treshiba, Levemir
9	GLP1 RA	Victoza, Byetta, Byudereon
10	SGLT2 Inhibitors	Suglat, Forxiga, Lusefi, Deberza, Apleway, Canaglu

Definition 2.2 *A full prescription $P(pid, did, s, e, m[], d[])$ is a tuple of pid patient id, did doctor id, s starting time, e end time, $m[]$ array of medicine name, and $d[]$ array of medicine dosages w.r.t. the medicine label. A prescription dataset is a sequence of prescriptions $[P_1, P_2, ..., P_n]$, where prescriptions are ordered by the starting time and duration. However, because we do not consider the switch of doctor events in further analyses, we also simplify the full prescription definition into a tuple of $P(pid, s, e, m[], d[])$. This simpler definition is used for further analyses.*

Table 2.3 presents an example of a prescription dataset for a patient from day 1 until day 1070. This example represents a progressive medication model of a patient from the actual dataset provided by Kyoto University Hospital.

Table 2.3: Full prescription dataset.

P_n	pid	did	s	e	m[]	d[]
P_1	7	1	1	56	A	d_a
P_2	7	1	64	134	A	d_a
P_3	7	1	191	256	A	d_a
P_4	7	1	257	340	A	d_a
P_5	7	1	347	375	A	d_a
P_6	7	1	376	390	B	d_b
P_7	7	1	397	407	A	d_b
P_8	7	1	406	420	A	d_a
P_9	7	1	406	435	B	d_b
P_{10}	7	1	421	443	A	d_a
P_{11}	7	1	436	443	A	d_a
P_{12}	7	1	450	481	A,B	d_a,d_b
P_{13}	7	2	482	570	A,B	d_a,d_b
P_{14}	7	2	630	690	A,B	d_a,d_b
P_{15}	7	2	691	778	A,B	d_a,d_b
P_{16}	7	3	749	820	C	d_c
P_{17}	7	2	779	840	A,B	d_a,d_b
P_{18}	7	3	821	900	C	d_c
P_{19}	7	4	841	900	A,B	d_a,d_b
P_{20}	7	4	901	998	A,B	d_a,d_b
P_{21}	7	4	901	998	C	d_c
P_{22}	7	4	950	960	A,B,C	d_a,d_b,d_c
P_{23}	7	4	955	967	D	d_d
P_{24}	7	4	968	975	A,B,D	d_a,d_b,d_d
P_{25}	7	5	976	998	D	d_d
P_{26}	7	5	1005	1070	A,B,D	d_a,d_b,d_d

2.3.1 Prescription Relation

The prescription relation represents the temporal relation between prescriptions in a time line. Figure 2.4 shows the possibility of temporal relations between consecutive prescriptions. For example, P_1 $before$ P_2, P_8 $starts$ P_9, and P_9 $overlaps$ P_{10}.

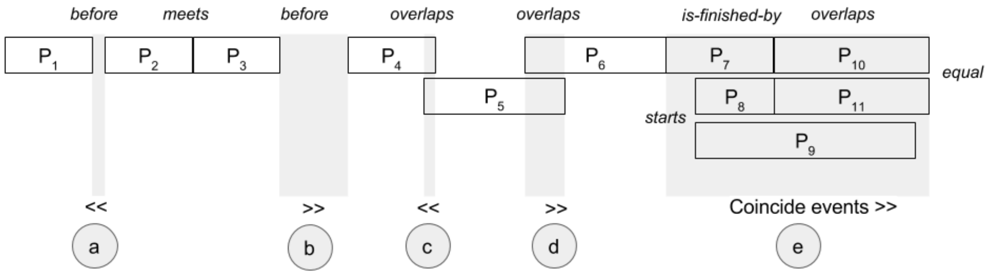


Figure 2.4: Allens temporal relations in aligned prescription sequence.

From Figure 2.4, event **a** and event **b** have the same relation (*before*). However, the gap on event **a** is very small compared to that on event **b**. It is similar with event **c** and event **d** (i.e., the overlap duration on event **c** is considerably smaller than on event **d**). In medication episode construction, such conditions may require different treatment. For example, the fact that a patient may come earlier or later than the appointment may cause short gaps and overlaps and should be treated as a *meets* prescription. Meanwhile, a longer gap and overlap should be treated as it is. However, the maximal gap for medication episode construction will only influence prescription with gap (*before* relation). For overlapping prescriptions, irrespective of how small the overlaps are, the duration will be treated as overlaps. In this circumstance, the time error margin (ϵ) is suitable for assigning the prescription relation.

Definition 2.3 *Epsilon, ϵ, is user-specified threshold. Using epsilon, the time point relations of equal "=" and less than "<" become more flexible by $\pm\epsilon$. Given that t_1 and t_2 are two time points, the following equations are true :*

$$t_1 =_\epsilon t_2 \leftrightarrow |\ t_1 \text{ - } t_2\ | \leq \epsilon$$

$$t_1 <_\epsilon t_2 \leftrightarrow 0 < t_2 \text{ - } t_1 > \epsilon$$

Example 2.1 *Based on Figure 2.5, if we use the notion of ϵ, then the prescription relations in Figure 2.5(a) and (c), which were previously A $before$ B and A overlaps B, will be A meets B. For Figure 2.5 (b) and (d), the prescription relation will remain the same.*

To demonstrate the prescription relations from the raw data, Figure 2.6 shows the prescription diagram of Table 2.3. To simplify, we do not display the dosage

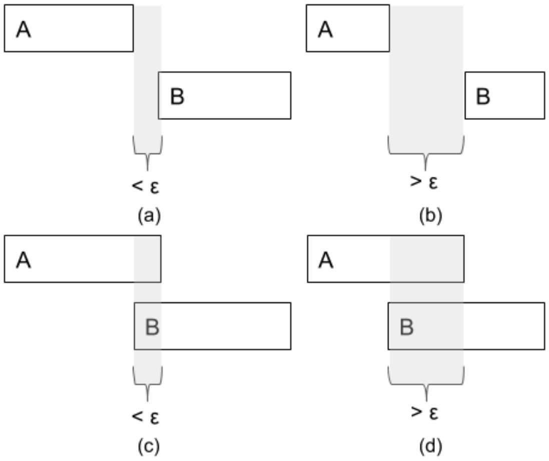


Figure 2.5: Time error margin (ϵ).

information in the prescription diagram. The x-axis shows the time per 30 days. Figure 2.6(a) shows the aligned prescriptions duration based on their start time, end time and medicine label. As shown in Figure 2.6(a), a prescription can have a short duration, and it occasionally overlaps another prescription or has a gap. As previously mentioned, this situation may occur because a patient may come earlier or later than the appointment with the physician. Another event that we observe is that many of the prescriptions continue the previous medication. Furthermore, the prescriptions begin overlapping each other when the prescription is modified by the physician. For example, with the transition from medicine A to dual therapy AB around $time = 13$, we have overlapping prescription between medicines A and B. Another example is shown when the medication modified from dual therapy AB to ABC around $time = 24$ and when the medication is switched from ABC to ABD around $time = 32$.

Example 2.2 *Using P_1, P_2, and P_3 from Table I and $\epsilon = 14$ days, we have two prescription relations, as follows :*

$$|P_2.s - P_1.e| < \epsilon \Rightarrow P_1 \ meets \ P_2$$

$$|P_3.s - P_2.e| > \epsilon \Rightarrow P_3 \ before \ P_2$$

2.3.2 Medication Episode

To reconstruct a continuous medication episode as in Figure 2.6(b), we use Allen's temporal relations relaxed by the notion of ϵ, as shown in Figure 2.1. Our main

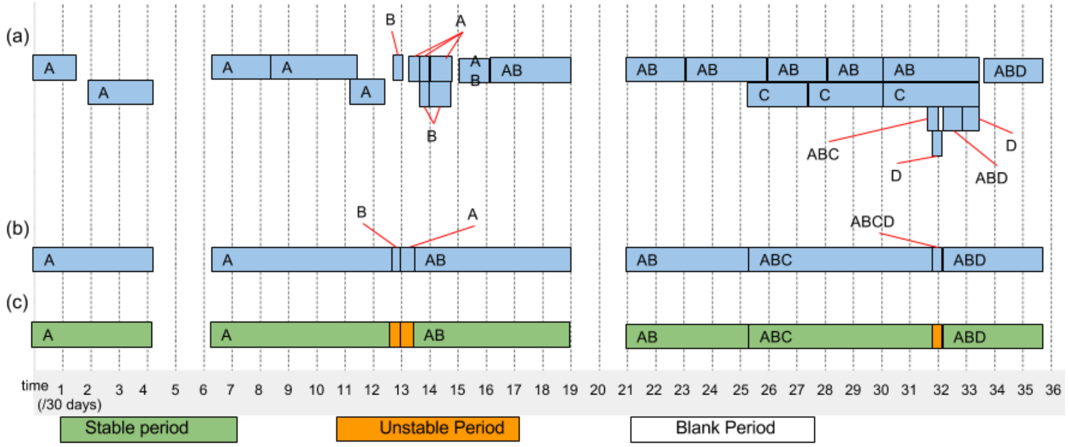


Figure 2.6: Physician's prescription diagram as listed in Table 2.3.

idea is to concatenate the same prescriptions with a *meets* relation to assemble a medication episode. Moreover, for any two prescriptions P_i and P_j , we aggregate types of relations, P_i *is−finished−by* P_j , P_i *contains* Pj , and P_j *starts* P_i ; we denote them as P_i *contains* P_j. The *contains* relation in clinical condition may occur during hospitalization conditions, where a patient should take the medicine from the hospital, and in such cases, the physician is adjusting the medications based on the patient's condition. For equal prescription, that is, prescription with the same start time and end time, we merge the prescriptions. A more detail on the rules of medication episode construction explained in the previous publication [26].

Definition 2.4 *A medication episode ME is a concatenation of meets prescriptions that have the same medicine label and dosage. ME shows the period of time when the physician does not change the prescription. The ME dataset is $\boldsymbol{ME}$ = $\{$ ME_1,ME_2,..,ME_n $\}$, where n is the total number of ME in the patient medical history and ME_n is ordered based on the starting time and end time.*

Example 2.3 *From Table 2.3, P_1 meets P_2, $P_1.m[] = P_2.m[]$, and $P_1.d[] = P_2.d[]$. Hence, P_1 and P_2 are concatenated into a single ME.*

Recalling Figure 2.6(b), we have nine medication episodes after reconstructing the prescriptions. By using the reconstruction results, we are able to distinguish the medication episode types and identify the stable periods as shown in 2.6(c). An unstable period represents short medication changes that may occur when the physician attempts to adjust the medication or because of hospitalization. Thus, to find the effective medication pattern in the long term, we consider a stable period to be essential for further analysis.

Definition 2.5 *Threshold δ is the minimum period, in days, for the physician to see the medications effectiveness.*

Definition 2.6 *A stable period SP is a medication episode in which the duration is at least equal to δ days. It is defined as $SP = \{SP \in \boldsymbol{ME} | SP.e - SP.s \geq \delta\}$.*

In addition, we define several other periods of time as the following.

Definition 2.7 *A trial/short period TP is a medication episode whose duration is less than the threshold δ days. It is defined as $TP = \{TP \in \boldsymbol{ME} | TP.e - TP.s < \delta\}$.*

Definition 2.8 *An Unstable period UP, is a single TP or an aggregation of consecutive TPs.*

Definition 2.9 *A Blank period BP, is a period of time when there was no medication recorded in the medical history after ϵ days.*

2.3.3 Medication Transition Events

After the stable period identification, we attain the SP sequence. From Figure 2.6(c), we have an SP sequence of $A \rightarrow A \rightarrow AB \rightarrow AB \rightarrow ABC \rightarrow ABD$. Medication transition events occurred in the point between $A \rightarrow AB$, $AB \rightarrow ABC$, and $ABC \rightarrow ABD$. We list five medication transition events as follows:

- *Add*, is when new medicine(s) added to the previous medication.
- *Stop*, is when previous medicine(s) stopped from the previous medication.
- *Switch*, is when new medicine(s) added and previous medicine(s) stopped.
- *Increasing*, is when the dosage of medicine(s) increased.
- *Decreasing*, is when the dosage of medicine(s) decreased.

In addition, we consider events, such as $A \rightarrow A$ and $AB \rightarrow AB$, as *continue* events.

Example 2.4 *Based on Figure 2.6(c), we consider $A \rightarrow AB$ and $AB \rightarrow ABC$ as add events and $ABC \rightarrow ABD$ as switch event.*

2.3.4 Rules of Construction

In this section, we give the detail rules of construction along with the appropriate assumptions. In this study, we consider five rules construction. However, the constructions are not limited to these rules and can be easily customized depends on the assumptions used on how handling the prescriptions relation.

The prescription relation described in Section 2.3.1 is then utilized for assigning the rules of construction. Each of these prescription relations will have specific treatment. In this section, we will list the rules of reconstruction and use Figure 2.7 to illustrate the implementation of the rules. The construction rules are developed based on the clinical assumption on drug exposure of diabetes treatment.

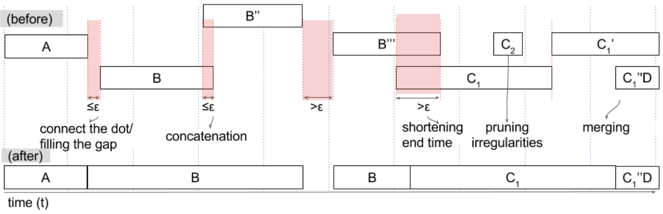


Figure 2.7: Implementation of the construction rules.

Rule of filling the gap/connecting the dots. We consider *meets* prescriptions to be connected to one time point. For this case, we assume that it is possible that a patient came to the physician earlier or later than the reserved schedule. Figure 2.7 shows prescriptions that have small gap with different medications. The end time of the first prescription and the start time of the next prescription will be connected at one point.

Rule of concatenation. For prescriptions that continue the previous medication and have a *meets* relation, we concatenate them to form a medication episode. As illustrated in Figure 2.7, the end time will shift to the end time of the next prescription.

Rule of shortening the end time. When the medicine is changed and creates an *overlap* condition, we assumed that the patient will take the current prescriptions. Therefore, we shorten the previous prescription's end time, as shown in Figure 2.7. This rule applies for changing the medicine regardless of the medicine types.

Rule of pruning irregularity. As demonstrated in Figure 2.7 for *contains* relation, we consider pruning the *contained* medicine that has the same medicine type. This is because from a medical point of view, it is uncommon for physicians to prescribe different medicines that have the same type in *contains* relation. However if the *contained* medicine belongs to another type of medicine, we leave it as it is.

Rule of merging. Occasionally, as the patient's condition progresses, the physician will attempt to adjust the medicine, which will lead to a *contain* relation. Another possibility for a *contain* relation is hospitalization. The *merging* rule will add the dosage of the same medicine and merge medicines of different types. However, different medicines of the same type will be pruned. The *merging* rule is also applied in the *overlaps* duration of overlapping prescriptions and *equal* prescriptions. We do not add the overlap duration at the end of the subsequent prescription as suggested by the second method of [18] to avoid a ripple effect. This ripple effect can be clearly seen in Figure 2.8. Using ϵ value of 14 days, originally A" and A"' has $before$ relation. Because of the ripple effect, this relation is changed to *meets* relation. Hence, the second method introduced by [18] is not an option for our case.

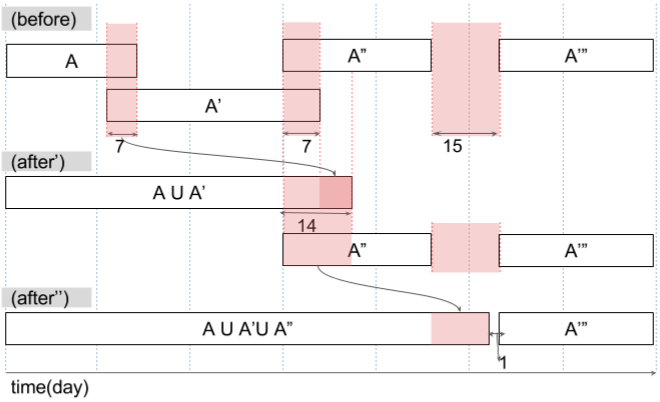


Figure 2.8: Ripple effect.

2.4 Experiment

The investigation in this section focuses on three aspects which support our main contributions, as follows: the nature of multitherapy dataset, the significance of ϵ in connecting successive prescription, and the stable periods generation influenced by ϵ value variation.

2.4.1 Dataset

We use an anonymized dataset provided by Kyoto University Hospital along with the approval from The Ethics Review Board of The Medical School of Kyoto University. The dataset is the prescription registry of T2DM patient's hospital prescription. The prescription is extracted between September 2000 and August 2015 for the medicines listed in Table 2.2.

We exclude patients with medicine types 8 and 9 because there is no information about the duration. We are left with 227,269 records(154,598 prescriptions out of 6,573 patients).

2.4.2 Result

First, to show the nature of multitherapy dataset, we extracted the numbers of prescription relations based on Allen's relation. Figure 2.9 shows the number of each prescription relation: $before$, $meets$, $overlaps$, $is - finished - by$ (isfby), $contains$, $starts$, and $equal$. Figure 2.9 shows that the $meets$ relation dominates the number of prescription relations followed by $before$, $overlaps$, $starts$, $contains$, $is - finished - by$ and $equal$. $Equal$ prescriptions represent prescriptions with the same time range given by a different physician as defined in the full prescription.

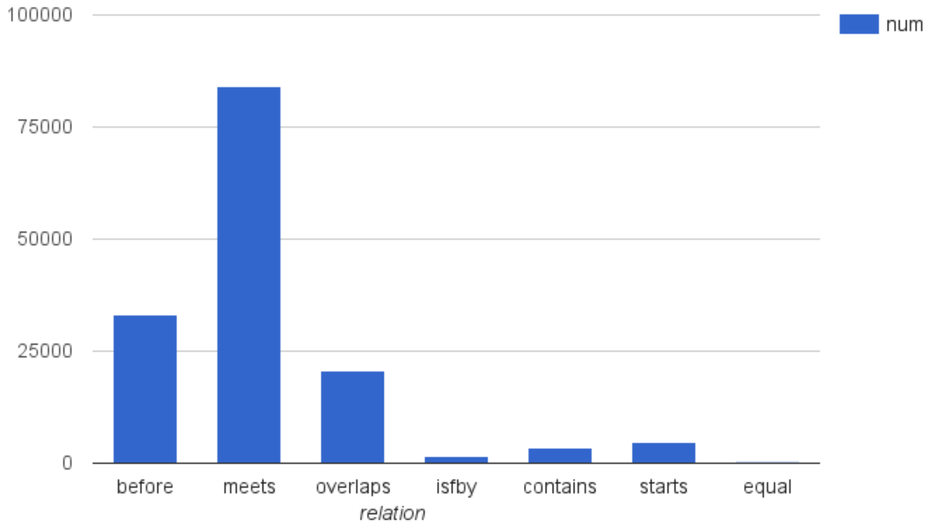


Figure 2.9: Number of each prescription relation extracted based on the fixed Allen's relation.

Second, to be able to observe how much of the *before* and *overlaps* portions that are influence by ϵ, we investigated the numbers of prescriptions with *before* and *overlaps* relations that has a gap and overlap $\leq$ than ϵ value variation. In this investigation we varied the ϵ value to 7, 14, and 21 days. This choice is based on our initial assumptions that a patient may come earlier or later than the appointment, and it is presented in Figure 2.10 and Figure 2.11. From the Figure 2.10, we are able to observe that it is more than 30% of the *before* prescriptions were influenced using the smallest value of ϵ (7 days). And the numbers of influenced prescriptions is increasing as the ϵ value increased. From Figure 2.11, more than 80% of the *overlaps* prescriptions are influenced by the ϵ value.

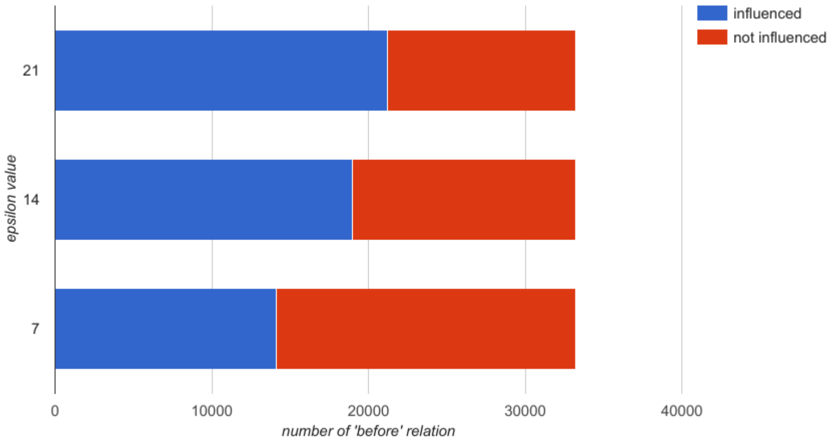


Figure 2.10: Prescription number with *before* relations.

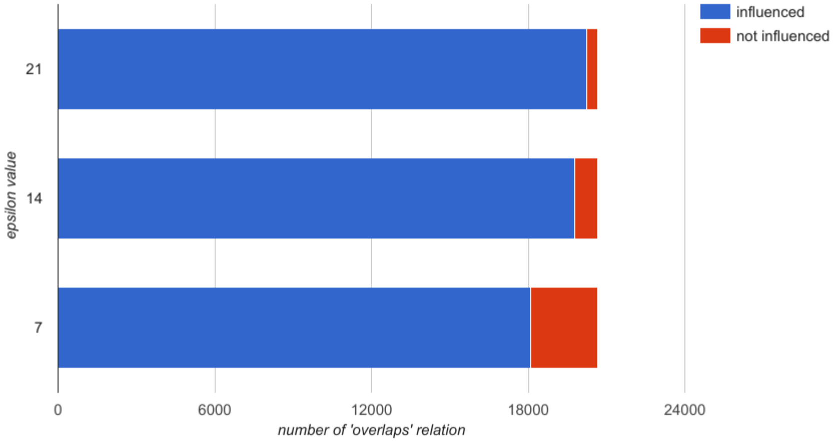


Figure 2.11: Prescription number with *overlaps* relations.

We also investigated the number of generated stable period that have certain duration upon variation in the value of ϵ value variation, as presented in Figure 2.12. The lines display the accumulated number of stable periods. The x-axis represents the duration per 90 days. The blue line represents the number of stable periods generated by the construction without ϵ (ϵ=0). The red, orange, and green lines are generated by the construction with ϵ value of 7 days, 14 days, and 21 days respectively. From the figure, we are able to observe that the number of stable periods with duration less than 200 days is sharply decreasing when we use the notion of ϵ (blue, red, and yellow lines). The blue line has higher number of short stable periods compared to other lines (under 200 days). However, the production of longer duration stable period without ϵ decreases (+- above 300 days).

To be able to observe the discrepancy from each line for duration more than 300 days, we also generated the log scale of Figure 2.12, which is displayed on Figure 2.13. Figure 2.13 shows that the blue line falls (construction without ϵ) under the other lines for stable periods with a duration of more than 300 days. Another observation from both figures is that the stable periods generated using ϵ have only slight differences.

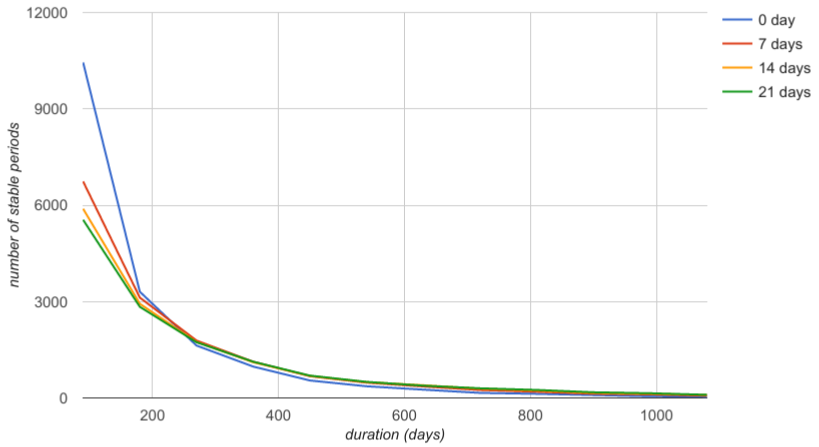


Figure 2.12: Number of stable periods with length duration.

A deeper observation on the number of stable period sequence from the constructed medication episode is demonstrated in Figure 2.14. In this figure, the disparity between the ϵ value selection is clearly observable. The number of stable period sequence is decreasing with higher ϵ value. Further observation between the number of continue pattern and transition events in the stable period sequence

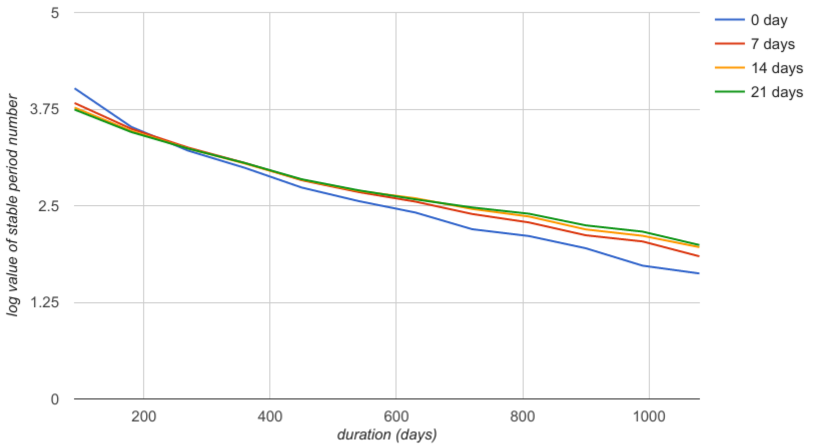


Figure 2.13: Log scale from Figure 2.12.

is shown in Figure 2.15. Figure 2.15 shows that the number of continue events are significantly decreased with the increased of ϵ value. In contrast, the number of transition events available to observed is increased as the ϵ value increased. To be noted, that the total number between continue events and transition events is not the same as the number of stable period sequence. This differences is because in one transition point there can be more than one transition events.

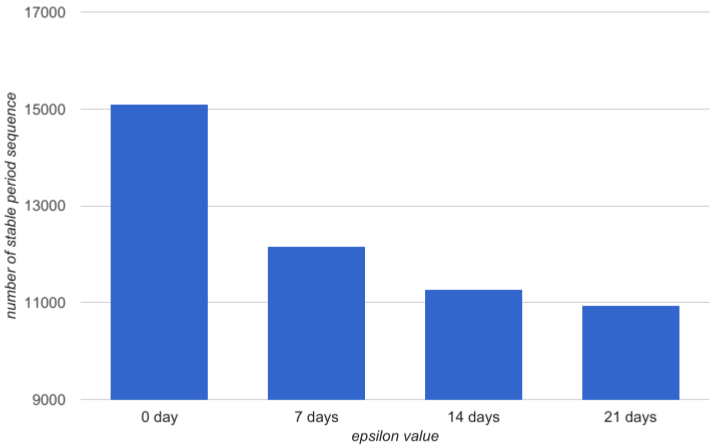


Figure 2.14: Number of stable period sequence based on the ϵ value.

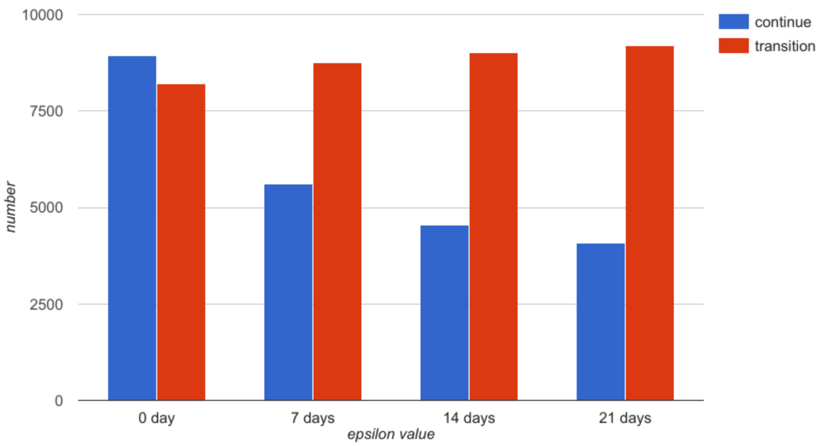


Figure 2.15: Number of continue event and transition event based on the ϵ value.

2.5 Discussion

In this section, we would like to discuss the results from the previous section. As shown in Figure 2.9, the prescriptions relations in a multitherapy dataset cover all Allen's relations. The high number of the *meets* relation shows that it is common for chronics patients prescriptions to be connected to each other caused by either the patient behavior (to come on schedule) or the patient condition that required the physician to customized the medication in series of connected prescriptions. When the *meets* prescriptions are having the same medication, they will be concatenated and form longer medication episode. Hence, the medication episode construction is still able to generate longer medication episode even without using ϵ. However, the numbers of $before$ and $overlaps$ relations are also significant in a multitherapy dataset. Hence, the incidents that patients come earlier or later than the schedule is also common. This patient behavior of coming earlier than the schedule may occur when a patient is unable to come on time because of other reasons and then decided to come earlier to renew the prescription. In this condition, the overlapping duration is considered short. However, in chronic patients, it is possible for patients to have changes of conditions which cause them to come earlier than the schedule. This condition may have longer overlapping duration. As for the patient behavior of coming later than the schedule may occur when a patient decided to come later because the patient unable to come on schedule. In this condition, the gap duration is usually short duration because for chronic patients it is important for them to take the medication. Another condition may cause the gaps, that is because the patient come to visit private clinic or other health care provider. This condition is usually caused longer gap. Hence, we need to be able to identify which of these $before$ and *overlaps* prescriptions have short gap or overlap duration. Thus, we need to take consideration for such incidents in the construction and to develop the rules for the medication episode.

As shown in the results in Figure 2.10 and Figure 2.11, the proportion of prescriptions with $before$ and *overlaps* relations that were influenced by ϵ are significant. Using the notion of ϵ, the construction of medication episode were able to identify prescriptions with short $before$ and *overlaps* relations from the longer ones, and then transformed them as prescriptions with *meets* relation. Compared to previous method in [18] that used maximal allowed gap, the method only influenced prescriptions with a $before$ relation while the prescriptions with *overlaps* relation remained having the same relation despite their short overlaps duration. From Figure 2.11, we are able to observe that there are prescriptions with short and longer overlaps duration. If using the method in [18], the size variation of *overlaps* will not be addressed because it used the fixed temporal relation. In this case, short overlap duration will not be identified, and then treated as overlap and merged in case of the medications from both successive prescriptions are different, which later will produced more short unstable periods.

In addition, there are considerably low numbers of $is-finished-by$, *contains*, *starts* and *equal* relations from Figure 2.9. Even though the low numbers, these prescription relations have significant meaning because they represent chronic patient's incidents that occasionally occur when there are temporarily abrupt changes in patient's condition or when the disease is progressing and needs to be managed by the physician. These incidents are reflected in the observation from previous example (Figure 2.6), which is when the patient's condition changes, medication transition events occur with prescriptions with a *contains* ($is-finished-by$, *starts*, *contains*) relation. Therefore, to retain the medication transitions information, we need to address those temporal relations.

Regarding other technical aspects of the medication episode construction, [18] addressed the effect of adding or not adding the duration of overlaps at the end of the predicted episode on the median length and the patient proportion number based on the length of the episode with a variation value in the value of the maximal allowed gap. As an addition to the discussion of our study, we would like to add that using the notion of ϵ ($\epsilon > 0$) also has an effect on the number of stable periods. This is shown in Figure 2.12. Without ϵ, the medication episode construction still generated stable periods. The stable periods generated without ϵ were produced in a very significant number for durations of less than 200 days. These results is in contrast with stable periods generated with using ϵ, which is shown with the large gap between the blue line and other lines in shorter duration (less than 200 days). This data behavior may be cause by many of stable periods generated without ϵ are merely a continuation of the previous stable periods but separated by a short blank period (gap). The short blank periods are created because the construction without ϵ will not be able to identify short gaps, which may be caused by the patient come a little later than the schedule. Therefore, the usage of ϵ is significant to avoid such condition. Moreover, based on the generated stable periods as shown in Figure 2.13, the higher ϵ value selection will have more performance in producing stable periods with longer durations (more than 300 days). This result is because the construction using ϵ will be able to connect prescriptions separated by short gaps, which will produce stable periods with longer duration. Hence, ϵ is essential for producing longer expression out of prescriptions. Furthermore, this result shows that the number of stable period transition available for further analysis (search space) is affected by the selection of ϵ value, and the search space size influence the cost of data driven analysis [1].

Moreover, from Figure 2.14, we are able to observe that the selection of ϵ value also influences the number of stable period sequence in each patient. Further observation presented in Figure 2.15 shows that the number continue event in the stable period sequence is sharply decreased. This result confirm the previous statement that many of stable periods are merely continuation from the previous one, which are then concatenated by the ϵ. In contrast, the number of transition events are increasing which shows that there are Unstable Periods connected by ϵ which then construct a stable period and more medication transition events able to observe. Regarding to this result, a study by [32] introduced a horizontal support value, which is the number of instances of the pattern found in an entity (e.g., a single patient medical record). Hence, in analyses based on the horizontal support value, the frequency outcome will show a high frequency with a lower ϵ value for continue events. Conversely, a lower frequency will be the outcome for higher ϵ values.

From a medical perspective, currently, clinicians asses the chronic diseases medication from the medication history in the form of prescription datasets, which are difficult to use particularly for longitudinal analysis because of their natures (e.g., short durations, fragmented and repetitive). Medication episode construction that enables expressing longer durations of medication history will provide a new means for obtaining long term clinical finding. As in diabetes, medication effect is commonly assessed in longer duration observation windows. Hence, a longer duration of stable periods is more relevant in drug utilization or pharmacoepidemiology studies compared to short prescriptions. For example, in studies that takes duration as an essential factor to investigate: the drug exposure and drug survival analyses (i.e., studies which assumes that a drug surviving longer in treatment will be one that is safer and/or more effective).

Finally, the dataset is originated from patients who went to Kyoto University Hospital (not an integrated dataset from multiple hospitals). Therefore, we would like to add more annotation about the complexity of longitudinal and multitherapy prescription analyses on our dataset. Considering multitherapy, there are *equal*, *overlaps* or *contains* prescriptions. We cannot be sure whether the physician is attempting to enhance or change the medicine dosage or even if there was hospitalization because in the case of hospitalization, the patient is taking medicine provided by the hospital only. The current assumption used in our rules is that coinciding medication will be merge (i.e., different medicines with same medicine type will be pruned as defined in the rule of *merging*). However, such conditions happen usually in a short time (less than 3 months). Hence, for longitudinal analyses, we concerned with the medication transition events between stable periods rather than the unstable period.

2.6 Conclusion

This chapter studies the data preparation for retrospective database analysis for observing medication transition events. Best to our knowledge, there has no framework of medication episode construction that incorporated all possible Allens temporal relation for multitherapy dataset. By accommodating Allen's relation in the ruled based construction, we are able to preserve prescriptions information in a multitherapy dataset which will be missing otherwise. Furthermore, the usage of ϵ in expressing Allen's relations is significant in reducing repetitive medication episode, constructing higher numbers of longer medication episodes and enabling more medication transition events available to be observed.

This is important for the longitudinal analysis of chronic diseases, particularly to observe the strategic actions by the physician to achieve ideal condition for the patients. Compared to a previous study, [18] emphasized the discussion on the gap influence on the median length of the episode and patient number. In addition, we completed the discussion from a technical perspective that selection of the epsilon value influences the generation of stable periods not only with respect to the duration but also the number of continue events and the medication transition events. Hence, our investigation of the selection of the ϵ affects the measurement of further analysis results significance.

CHAPTER 3

Singleton Mining of Medication Strategy

3.1 Motivation

Frequent sequential pattern (FSP) mining was developed for application to solve the market basket, that is to study customer behavior [1]. FSP mining utilizes a sequence database of customer transactions to look for frequent sequences of itemset pattern that have a support value greater than a threshold value defined by the user. The support value is counted based on the occurrences of the pattern in the sequence database.

FSP approach is a practical method of finding frequent sequence events and has been applied in many fields. For example in the medical field, in the area of postcare, [6] reports an application regarding patients in the follow-up of liver transplantation, [5] reports a study on the ICU patients to detect recent events, and in the field of pharmacovigilance, [21] presents an investigation of adverse drug reaction using FSP mining.

Recently, in the area of chronic diseases such as type 2 diabetes, the FSP approach is also used to mine electronic medical records (EMR). Diabetes type 2 is a typical chronic disease associated with its high numbers of risk factors and parameters that affect the therapy. FSP mining enables users to yield new insights from existing therapy data regarding the medication trend or guideline adherence. In [53], new clinical condition of diabetes type 2 patients were found compared to the medical guidelines and [55] uses FSP mining to predict the next prescribed medicine in monotherapeutically medicated diabetes type 2 patients.

Our interest is in observing the medication transition events to learn about physician's strategy in managing the illness. Current method demonstrate FSP mining on monotherapy [55], whereas in reality, physician's prescriptions may vary from monotherapy to multitherapy. Furthermore, in the case of diabetes, the selection of pharmacotherapy is considered essential [51]. The appropriate combination of medications should be selected in accordance with the patient conditions. Thus, using Apriori [1] as it is may result in patterns that do not represent the actual medications. This is because Apriori-based FSP mining finds frequent sequence of items set that can be a partial item sets. In addition, Apriori-based algorithm considers sequences that may not occur consecutively. By contrast, we consider the consecutive order of the sequence.

In order to solve the type of problem considered here, we introduced singleton mining in [26]. The difference between singleton mining and Apriori-based FSP mining is as shown as follows: In Apriori algorithm, the sequence of $\langle Sulfonylurea\rangle \rightarrow \langle DPP4-inhibitor\rangle$ is contained in $\langle Sulfonylurea, \alpha GI\rangle \rightarrow \langle \alpha GI, Thiazolidinediones\rangle \rightarrow \langle \alpha GI, DPP4-inhibitor\rangle$; whereas in singleton mining, it is not considered as an instance of sequence.

In this section, we present a methods-comparison study between singleton mining and Apriori-based FSP mining to investigate the pattern results produced by both methods, in order to survey the characteristic of both methods' results. In addition, we conduct a confirmatory experiment to answer the clinical physicians research question. As recently increased hypoglycemia in aged patients, physicians try to diminish Sulfonylurea (SU) usage. However, once a treatment is start with SU, it is hard to replace the medication with other medication. The clinical physician want to understand whether the new released medication has impact towards the doctor behavior in diminishing the usage of SU. Therefore, we conduct a confirmatory experiment to investigate the underlying impression of the clinical physician toward the new released drug.

3.2 Related Works

3.2.1 Frequent Sequence Pattern Mining

Our work is a case specific extension of frequent sequence pattern mining, first introduced by [1]. The problem of frequent pattern mining could be explained as follows [29]. Let $I = \{i_1, i_2, ..., i_m\}$ be a set of m distinct items. Items are ordered by a total order on I. An event (also called itemset) of size L is a non empty set of L items from I, which is sorted in increasing order. A sequence α of length n is an ordered list of n events $\alpha_1, \alpha_2 ..., \alpha_n$ denoted as $\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_n$, which is ordered based on the timestamp. A sequence database D is composed of sequences, where each sequence has a unique sequence identifier (sid). A sequence $sa = \alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_n$ is contained in (subsequence of) another sequence $sb = \beta_1 \rightarrow \beta_2 \rightarrow ... \rightarrow \beta_m$ if and only if there exist integers $1 \leq i_1 < i_2 < ... < i_n \leq m$ such that $\alpha_1 \subseteq \beta_{i_1}, \alpha_2 \subseteq \beta_{i_2}, ..., \alpha_n \subseteq \beta_{i_n}$. We consider medication types in Table 2.2 as the list of items and Table 3.1 as the sequence database. The patient id pid as the sequence id. An event is a medication episode, which shows a combination of medication(s) that is given to the patient in a period of time. A sequence is a set of ordered medication episodes based on their timestamps.

Example 3.1 *Table 3.1 presents the medication type transitions of the stable periods from six patients' medical record. The characters inside the bracket denote the medicine type(s) and the arrow represents the transition. From Table 3.1, for patient id 2, we have five medication episodes that are the combination of medication type as follow:* $\langle \alpha GI, Big \rangle$, $\langle \alpha GI, THZ \rangle$, $\langle SU, \alpha GI, THZ \rangle$, $\langle SU, Big \rangle$, *and* $\langle SU, Big, DPP4i \rangle$. *The sequence* $\langle \alpha GI, Big \rangle \rightarrow \langle \alpha GI \rangle \rightarrow \langle SU, Big \rangle$ *is contained in patient id 2 sequence. because* $\langle \alpha GI, Big \rangle \subseteq \langle \alpha GI, Big \rangle$, $\langle \alpha GI \rangle \subseteq \langle \alpha GI, THZ \rangle$ *and* $\langle SU, Big \rangle \subseteq \langle SU, Big \rangle$. *However, the sequence* $\langle SU \rangle \rightarrow \langle Big \rangle$ *is not contained in* $\langle SU, Big \rangle$.

The task in frequent sequential pattern mining is to find all sequence patterns p, which are frequent in a database D if p is contained in at least a certain percentage (support) of sequences of D. The problem in frequent sequential pattern mining is that the number of possible pattern candidate is exponential. Therefore, to explore all the possibility could consume a high computation cost. Apriori principle is used to solve the problem. Apriori principle, which is that

Table 3.1: An example of a sequence dataset of medication type transition.

pid	medType Transition
1	$\langle SU, \alpha GI\rangle \rightarrow \langle \alpha GI, THZ\rangle \rightarrow \langle \alpha GI, DPP4i\rangle$
2	$\langle \alpha GI, Big\rangle \rightarrow \langle \alpha GI, THZ\rangle \rightarrow \langle SU, \alpha GI, THZ\rangle \rightarrow \langle SU, Big\rangle \rightarrow \langle SU, Big, DPP4i\rangle$
3	$\langle SU\rangle \rightarrow \langle SU, Big\rangle \rightarrow \langle SU, Big, DPP4i\rangle \rightarrow \langle Big, DPP4i\rangle$
4	$\langle SU\rangle \rightarrow \langle SU, Big\rangle \rightarrow \langle Big, DPP4i\rangle$
5	$\langle THZ\rangle \rightarrow \langle Big, THZ\rangle \rightarrow \langle Big\rangle$
6	$\langle SU\rangle \rightarrow \langle SU, Big\rangle \rightarrow \langle SU, \alpha GI, Big\rangle \rightarrow \langle SU, Big, DPP4i\rangle$

if a sequence is frequent then the itemsets, which constructed it, must be also frequent. Using the Apriori principle, itemsets that are not frequent and their super sequences (i.e., sequences that contain the itemset) will be pruned.

3.2.2 Negative Pattern Mining

Pattern mining described in previous section were developed to discover positive sequential patterns from the database. Positive sequential patterns consider only the occurrences of itemsets in a sequence. However, in practice, the absence of an itemset in a sequence may imply valuable information [30]. For example, a clinician may want to understand how a certain medication is diminished to use. As mentioned in previous section that Sulfonylurea (SU) may increase the risk of hypoglycemia in aged patients. And since then, clinicians try to avoid to start the medication using SU. Hence, negative pattern mining has potential to give valuable information about this phenomenon.

Negative pattern mining task is to find frequent pattern with a constraint that the itemset not containing certain item. A positive sequence is denoted by $\langle \alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_n\rangle$ and a negative sequence is denoted by $\langle \alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \neg\alpha_n\rangle$, where $\neg\alpha_n$ represents the absence of itemset α_n.

Example 3.2 *From Table 3.1, for patient id 1, we have three medication episodes as follow:* $\langle SU, \alpha GI\rangle, \langle \alpha GI, THZ\rangle$, *and* $\langle \alpha GI, DPP4i\rangle$*. The sequence* $\langle SU, \alpha GI\rangle \rightarrow \langle \alpha GI, THZ\rangle \rightarrow \langle \alpha GI, DPP4i\rangle$ *is considered to support a negative pattern with the following pattern :* $\langle SU\rangle \rightarrow \langle \neg SU\rangle$*.*

3.2.3 Similarity Measurement

To compare the result set of our proposed method and the conventional method, we use similarity measurement for comparing both result ranked lists. In this section, we discuss two methods for comparing ranked lists.

1. Overlap: A common similarity function is proposed to compute the ratio between the number of rules from one rule set that occur in another rule set [4]. Suppose that there are two rule sets R_1 and R_2. The overlapping of rules between sets is a basic measure to investigate the common properties of rule sets. The overlapping ratio as similarity function between a pair of rule sets is typically defined as the following [4]:

   $Overlap(R_1, R_2) = |R_1 \cap R_2| / |R_1 \cup R_2|$

2. Kendall's Tau coefficient. This quantity is a statistic used to measure rank correlation (i.e., the similarity of the orderings of the data when ranked by each of two quantities[8]). Kendall's Tau represents the difference between the probability that the observed data will appear in the same order versus the probability that the observed data will not appear in the same order [52].

   The equation of Kendall's Tau is the ratio between of the difference and the sum of the numbers of the concordant pairs (C) and the discordant pairs(D), $\tau = (C - D)/(C + D)$[35]. A concordant pair is when the rank of the second variable is greater that the rank of the former variable. As a discordant pair is when the rank is equal to or less that the rank of the first variable [39].

3.3 Methodology

3.3.1 Singleton Mining

As stated in the previous section, we are interested in investigating the transition events between stable periods (SPs). Hence, we focus on itemset changes that are located next to each other. Before we describe the medication pattern definition, first herewith is the terms used in the following section: A full itemset is an itemset that is contained equally in at least one itemset of the sequence data. As for a partial itemset is an itemset that is a partial subset of itemset in at least one itemset of the sequence data. Consider a sequence $\langle x\rangle \rightarrow \langle xy\rangle$. According to the definition here, $\langle x\rangle$ is a full itemset and a partial itemset.

Example 3.3 *Using Table 3.1, an itemset of $\langle SU, \alpha GI\rangle$ is a full itemset because it is contained equally with the itemset of patient id 1 sequence data. An itemset of $\langle \alpha GI\rangle$ is a partial itemset because it is a partial subset in the itemset of patient id 1s sequence data. However, an itemset $\langle SU\rangle$ is not a partial itemset, because it is a full itemset based on the sequences of patient id 3,4, and 6.*

Further, we define a frequent medication pattern as repeated medication events found in the sequence data set with the cardinality equal to or greater than the minimum support value. We consider two types of medication pattern, as follows :

1. A singleton pattern is a pattern of a full itemset of SP. The singleton pattern may be stated based on the medicine name, medicine type, or medicine label. In the case of diabetes, it may take the form of monotherapy, dual therapy, triple therapy or more.

2. n-sequence pattern is a sequence pattern consisting n+1 adjacent singletons.

The support of the pattern p is calculated as the ratio of number of patients who exhibit the pattern, at least once in their longitudinal medical history, to the total number of patients, $Support(p)$ = number of patient with p/number of patient.

Example 3.4 *For example, the patient with pid 1 has medication transition from dual therapy with SU and αGI to dual therapy with medicine types SU and THZ and then followed by the subsequent transition to dual therapy with medicines types αGI and DPP4i. Hence, from Table 3.1, with a minimum support value of 0.2, we can find four singleton pattern as follows: $\langle SU\rangle, \langle SU, Big\rangle, \langle \alpha GI, THZ\rangle$, and $\langle SU, Big, DPP4i\rangle$. The 1-sequence patterns are as follows : $\langle SU\rangle\langle SU, Big\rangle$, and $\langle SU, Big\rangle\langle SU, Big, DPP4i\rangle$.*

This pattern definition is different from that used to generate Apriori-based FSP results. In Apriori, it considers that the occurrences of a subset of the singleton support the frequent sequence and the even though the sequence does not in a consecutive manner it will add the cardinality of the pattern.

Example 3.5 *For the data in Table 3.1 and a minimum support value of 0.2, Apriori-based FSP mining will indicate* $\langle Big, DPP4i\rangle$ *as a frequent pattern with support value of 0.5 and* $\langle Big, DPP4i\rangle \rightarrow \langle SU, Big, DPP4i\rangle$ *as another frequent pattern with a support value of 0.5. By contrast, when singleton mining is applied,* $\langle Big, DPP4i\rangle$ *is not considered as a frequent pattern because it is not a full itemset, where as for* $\langle SU, Big\rangle \rightarrow \langle SU, Big, DPP4i\rangle$ *pattern, the calculated support value is only 0.33 because the patient with pid 6 is not counted as supporting the pattern support.*

3.3.2 Singleton Pattern Mining with Flexible distance

First, we develop singleton mining to understand the physicians reasoning in changing the medication. However, clinician may also want to understand, for example, for treatments that start with certain medication, what kind of medication combination the patient will be having in the future. Therefore, we extend the singleton mining to have a more flexible distance. If singleton pattern is denoted as $\langle \alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_n\rangle$, singleton pattern with flexible distance is denoted by adding "+" as denoted $\langle \alpha_1 \rightarrow^+ \alpha_2 \rightarrow^+ ... \rightarrow^+ \alpha_n\rangle$. The "+" symbol denotes that the sequence(steps) can be one or more.

Example 3.6 *Using Table 3.1,* $\langle SU, Big\rangle \rightarrow \langle SU, Big, DPP4i\rangle$ *is a singleton pattern with support value of 0.33, while* $\langle SU, Big\rangle \rightarrow^+ \langle SU, Big, DPP4i\rangle$ *is a singleton pattern with flexible distance with support value of 0.5 because the patient with pid 6 is counted to be supporting the pattern.*

3.3.3 Hybrid Pattern Mining

The next requirement to give better understanding in learning the physician strategy is function to identify medication that is absence of certain medication. A clinical physician may want to understand if a treatment start with certain medication, what type of medication combination will the patient having in the future. This medication combination in question include the inclusion or the absence of certain medication. Hence, in addition, we combine the singleton pattern mining with the conventional pattern mining with constraint, in particular the negative pattern. We refer this hybrid pattern mining that is when we mine pattern that incorporate the notion of singleton (full itemset), subset (partial itemset) denoted by adding "*" in the itemset, and negative pattern.

Example 3.7 $\langle SU \rangle \rightarrow^{+} \langle \neg SU \rangle^{*}$*, is a pattern that its antecedent (the left side) is a singleton of SU and the consequent (the right side) is a negative subset that may contain any medication but SU. In addition, the distance is flexible. Using Table 3.1, the pattern* $\langle SU \rangle \rightarrow^{+} \langle \neg SU \rangle^{*}$ *is supported by value of 0.33*

3.4 Experiment

We conduct two types of experiment. The first one is a methods-comparison experiment and the second one is a confirmatory experiment. The comparative experiment aims to find frequent sequence patterns out of the data set using R-arulessequences library [10], which is the implementation of Apriori-based cSPADE algorithm [56], and our own method of singleton mining. Afterward, we inspect the result and compared the patterns set. As for the confirmatory experiment, we conduct experiment to investigate the clinical physician assumption towards the new released drug on the physician behavior in selecting medication for prescription.

3.4.1 The Dataset

We applied our methods on to a medical history of patients with diabetes, which is provided by Kyoto University Hospital along with the approval from the Ethics Review Board of The Medical School of Kyoto University. The dataset consists of prescription of diabetes type 2 outpatients from September 2000 until August 2015. First, we applied the medication episode construction framework and identified the stable periods from each patients. Then, we attain stable period sequences of 2461 patients in total. Table 3.2 shows the description of the studied population. The parameter age_0 is patient's age at the start date of the first stable period and age_n is patient's age at the end date of the last stable period. For the paramenter $Duration$, it is calculated by subtracting the start date of the first stable period with the end date of the last stable period and $1y$ is equal to 365 days. Furthermore, we divided the dataset into two parts that is before 2010 (10 years of dataset, 1781 patients) and after 2010 (5 years of dataset, 1898 patients). This partition is for the confirmatory experiment. The year of 2010 is

Table 3.2: Description of the studied population.

Variable	Condition	Number of Patients	Number of Patients
Age		age_0	age_n
	< 30	20	14
	30-40	63	32
	40-50	178	111
	50-60	507	238
	60-70	852	657
	> 70	841	1415
Duration			
	$\leq$ 1y	112	
	1y - 3y	560	
	3y - 5y	553	
	5y - 10y	788	
	$>$ 10y	448	
Gender			
	M	1508	
	F	953	

chosen because in that particular year, there was a new drug release. Table 3.3 shows the description of each population set.

3.4.2 Methods-comparison Experiment

The intention of the first experiment is to compare the result set of singleton mining and conventional mining. Figure 3.1 summarizes the experimental setting.

In this experiment, we conduct a similarity measurement between two ranking lists of our proposed method result set ($FP_{singleton}$) and the results of cSPADE (FP_{cspade}). However, this similarity function is not appropriate for measuring the similarity between the mining results of the methods considered here. This condition is because the Apriori algorithm does not require consecutive order of sequence or full itemset, the results of cSPADE (FP_{cspade}) will be a superset of the singleton mining results ($FP_{singleton}$). Moreover, because $FP_{singleton} \subset FP_{cspade}$. Therefore, the overlap score between $FP_{singleton}$ and FP_{cspade} will be: $Overlap|FP_{singleton}, FP_{cspade}| = |FP_{singleton}|/|FP_{singleton} \cup FP_{cspade}|$. Hence, the

Table 3.3: Description of the population set (before and after 2010).

Variable	Condition	Before 2010	After 2010
Duration			
	$\leq$ 1y	186	113
	1y - 3y	489	440
	3y - 5y	418	543
	5y - 10y	623	793
	$>$ 10y	65	9
Gender			
	M	1072	1167
	F	709	731

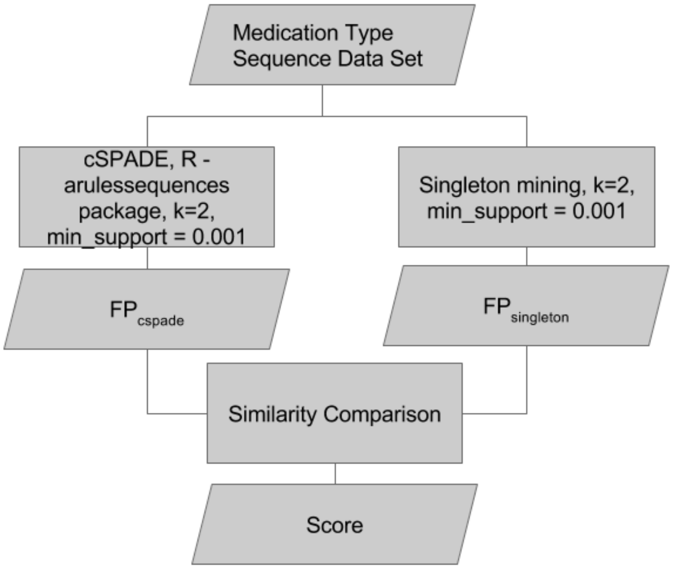

Medication Type Sequence Data Set
cSPADE, R - arulessequences package, k=2, min_support = 0.001
Singleton mining, k=2, min_support = 0.001
FP cspade
FP singleton
Similarity Comparison
Score

Figure 3.1: Experiment setting.

overlap score will always be the ratio of $FP_{singleton}$ and FP_{cspade}. We obtain value of 0.203 for the ratio value of $FP_{singleton}$ and FP_{cspade}.

Other similarity measure can be calculated using the Kendal rank correlation coefficient, commonly referred to as Kendall's Tau. Because Kendall's Tau is used to measure the order similarity between two rank sets of the same data but $FP_{singleton}$ and FP_{cspade} differ in numbers as $FP_{singleton} \subset FP_{cspade}$, there exist $FP_{singleton} \setminus FP_{cspade} = \{x \in FP_{cspade} \wedge x \notin FP_{singleton}\}$. For this case, we propose

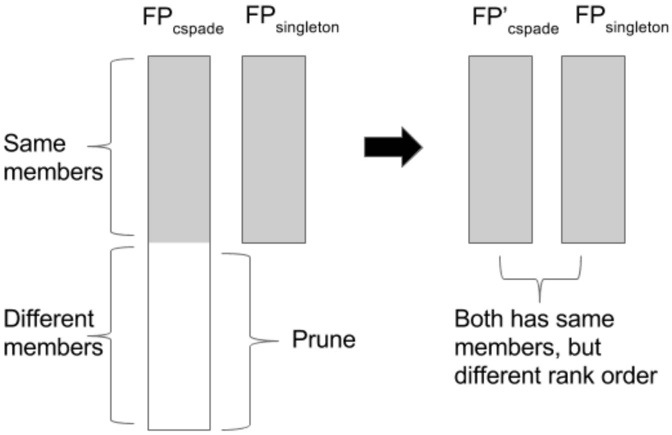


Figure 3.2: Kendall's Tau solution.

an alternative solution as follows. Prior to the calculating of the Kendall's Tau, we prune out the result members of FP_{cspade} that are the complement of $FP_{singleton}$, such that $FP'_{cspade} = FP_{cspade} \cap R_{singleton}$. In this way, we obtain two rank sets which have the same members ($FP_{singleton} = FP'_{cspade}$). Figure 3.2 illustrates the pruning of FP_{cspade}.

3.4.3 Results

Table 3.4 shows the top 20 results with the highest support values among the patterns identified by the singleton mining algorithm and the Apriori-based mining algorithm. From Table 3.4, we can observe that three patterns have the same rank, such as pattern no 1,2, and 5. However, the rankings of other patterns are jumbled. Some patterns appear in the same order in both sets of results; for example, the set of {⟨SU⟩, ⟨SU⟩ → ⟨SU⟩, ⟨SU,Big⟩, ⟨SU,Big⟩ → ⟨SU,Big⟩} has the same order in both $FP_{singleton}$ and FP_{cspade}. By contrast, the pattern $\langle \alpha GI \rangle$ comes after pattern ⟨SU,Big⟩ in $FP_{singleton}$, where as in FP_{cspade}, the ordering of these patterns is the opposite. Furthermore, patterns with the same base stem (antecedent) and consequent may have different ranks when ordered based on the base stem. For example the pattern with the base stem ⟨SU,Big⟩ and consequent ⟨SU,Big⟩ is ranked first in $FP_{singleton}$ but third in FP_{cspade} after ⟨SU,Big⟩ → ⟨SU⟩ and ⟨SU,Big⟩ → ⟨Big⟩.

We also consider the difference in the support values, as shown in Table 3.5. From Table 3.5, we observe that Apriori-based algorithm produces higher support

Table 3.4: Top 20 most frequent pattern sets identified through singleton mining and apriori-based mining.

No	$FP_{singleton}$	FP_{cspade}
1	$\langle \text{SU} \rangle$	$\langle \text{SU} \rangle$
2	$\langle \text{SU} \rangle \rightarrow \langle \text{SU} \rangle$	$\langle \text{SU} \rangle \rightarrow \langle \text{SU} \rangle$
3	$\langle \text{SU,Big} \rangle$	$\langle \text{DPP4i} \rangle$
4	$\langle \text{SU,}\alpha GI \rangle$	$\langle \text{Big} \rangle$
5	$\langle \alpha GI \rangle$	$\langle \alpha GI \rangle$
6	$\langle \text{SU,Big} \rangle \rightarrow \langle \text{SU,Big} \rangle$	$\langle \text{Big} \rangle \rightarrow \langle \text{Big} \rangle$
7	$\langle \text{Big} \rangle$	$\langle \text{SU} \rangle \rightarrow \langle \text{DPP4i} \rangle$
8	$\langle \text{SU,DPP4i} \rangle$	$\langle \alpha GI \rangle \rightarrow \langle \alpha GI \rangle$
9	$\langle \text{RaIS} \rangle$	$\langle \text{DPP4i} \rangle \rightarrow \langle \text{DPP4i} \rangle$
10	$\langle \text{DPP4i} \rangle$	$\langle \text{SU} \rangle \rightarrow \langle \text{Big} \rangle$
11	$\langle \text{SU} \rangle \rightarrow \langle \text{SU,Big} \rangle$	$\langle \text{SU,Big} \rangle$
12	$\langle \alpha \text{ GI} \rangle \rightarrow \langle \alpha \text{ GI} \rangle$	$\langle \text{Big} \rangle \rightarrow \langle \text{SU} \rangle$
13	$\langle \text{SU,}\alpha \text{ GI} \rangle \rightarrow \langle \text{SU,}\alpha \text{ GI} \rangle$	$\langle \text{SU} \rangle \rightarrow \langle \text{SU,Big} \rangle$
14	$\langle \text{SU,Big,DPP4i} \rangle$	$\langle \text{SU,DPP4i} \rangle$
15	$\langle Big \rangle \rightarrow \langle Big \rangle$	$\langle \text{SU,Big} \rangle \rightarrow \langle \text{SU} \rangle$
16	$\langle \text{SU,DPP4i} \rangle \rightarrow \langle \text{SU,DPP4i} \rangle$	$\langle \text{SU,Big} \rangle \rightarrow \langle \text{Big} \rangle$
17	$\langle \text{SU,}\alpha \text{ GI,Big} \rangle$	$\langle \text{Big} \rangle \rightarrow \langle \text{SU,Big} \rangle$
18	$\langle RaIS \rangle \rightarrow \langle RaIS \rangle$	$\langle \text{SU} \rangle \rightarrow \langle \text{SU,DPP4i} \rangle$
19	$\langle \text{Big,DPP4i} \rangle$	$\langle \text{SU,Big} \rangle \rightarrow \langle \text{SU,Big} \rangle$
20	$\langle \text{SU} \rangle \rightarrow \langle \text{SU,DPP4i} \rangle$	$\langle \alpha \text{ GI} \rangle \rightarrow \langle \text{SU} \rangle$

value than singleton mining and that there are 13 patterns have support difference > 0.1 out of 20 patterns.

Furthermore, we make a selection to the singleton mining and the Apriori-based mining with a criteria as follows : antecedent (the left side) is not equal to consequence and support is higher or equal to 0.005 and confidence (the right side) higher or equal to 0.01 and lift higher or equal to 1 and ordered by confidence in descending manner[*]. The results are ranked based on the confidence value and

[*] A confidence measure is a conditional probability of some event Y, given the occurrence of some other event X, $Confidence(p : X \rightarrow Y)$ = number of patient with p/number of patient with event X. In addition, a lift measure is a correlation measure that shows the rate of possibility of pattern that start with a certain event X will likely to change in to another

Table 3.5: Support value deviations for the 20 most frequent patterns in the singleton mining set that intersect Apriori-based mining set.

no	$Pattern$	$sup_{cspade} - sup_{singleton}$
1	⟨SU⟩	0.220864
2	⟨SU⟩ → ⟨SU⟩	0.284434
3	⟨SU,Big⟩	0.088019
4	⟨SU,α GI⟩	0.069275
5	⟨α GI⟩	0.244499
6	⟨SU,Big⟩ ⟨SU,Big⟩	0.10106
7	⟨Big⟩	0.312959
8	⟨SU,DPP4i⟩	0.13692
9	⟨RaIS⟩	0.072535
10	⟨DPP4i⟩	0.344744
11	⟨SU⟩ → ⟨SU,Big⟩	0.167075
12	⟨αGI⟩ → ⟨αGI⟩	0.229829
13	⟨SU,α GI⟩ → ⟨SU,α GI⟩	0.080685
14	⟨SU, Big, DPP4i⟩	0.022819
15	$\langle Big\rangle \rightarrow \langle Big\rangle$	0.318663
16	⟨SU, DPP4i⟩ → ⟨SU, DPP4i⟩	0.114914
17	⟨SU, αGI, Big⟩	0.015485
18	$\langle RaIS\rangle \rightarrow \langle RaIS\rangle$	0.067644
19	$\langle Big, DPP4i\rangle$	0.123064
20	$\langle SU\rangle \rightarrow \langle SU, DPP4i\rangle$	0.199674

Table 3.6 shows the Top 20 member on both result sets.

3.4.4 Confirmatory Experiment

The next experiment is driven by the clinical physician question on whether the new released medication has impact towards the doctor behavior in diminishing the usage of Sulfonylurea (SU). This question is because it is known that SU usage for long term will increase the risk of hypoglycemia in old age. In addition, hypoglycemia can cause dementia. Furthermore, with a new drug released in

certain event Y, $lift(p : X \rightarrow Y)$ = number of patient with $p/Support(X) * Support(Y)$.

Table 3.6: Top 20 most frequent patterns (FP) identified through singleton mining and coventional mining with a selection criteria.

No	$FP_{singleton}$	FP_{cspade}
1	$\langle SU, \alpha GI, Big, THZ\rangle \rightarrow \langle SU, \alpha GI, Big\rangle$	$\langle Big, THZ\rangle \rightarrow \langle Big\rangle$
2	$\langle SU, Big, THZ\rangle \rightarrow \langle SU, Big\rangle$	$\langle SU, \alpha GI, Big\rangle \rightarrow \langle SU\rangle$
3	$\langle Big, THZ\rangle \rightarrow \langle Big\rangle$	$\langle SU, Big, THZ\rangle \rightarrow \langle SU, Big\rangle$
4	$\langle RaIS, Big\rangle \rightarrow \langle SU, Big\rangle$	$\langle \alpha GI, Big, THZ\rangle \rightarrow \langle \alpha GI, Big\rangle$
5	$\langle SU\rangle \rightarrow \langle SU, Big\rangle$	$\langle SU, Big\rangle \rightarrow \langle SU\rangle$
6	$\langle SU, Big\rangle \rightarrow \langle SU, Big, DPP4i\rangle$	$\langle SU, \alpha GI, Big, THZ\rangle \rightarrow \langle SU, \alpha GI, Big\rangle$
7	$\langle SU, \alpha GI, THZ\rangle \rightarrow \langle SU, THZ\rangle$	$\langle SU, THZ\rangle \rightarrow \langle SU\rangle$
8	$\langle SU, \alpha GI, Big\rangle \rightarrow \langle SU, \alpha GI, Big, DPP4i\rangle$	$\langle RaIS, \alpha GI\rangle \rightarrow \langle \alpha GI\rangle$
9	$\langle RaIS, \alpha GI\rangle \rightarrow \langle RaIS, \alpha GI, Big\rangle$	$\langle RaIS, Big\rangle \rightarrow \langle Big\rangle$
10	$\langle RaIS, \alpha GI\rangle \rightarrow \langle \alpha GI\rangle$	$\langle SU, \alpha GI\rangle \rightarrow \langle SU\rangle$
11	$\langle Big\rangle \rightarrow \langle Big, DPP4i\rangle$	$\langle \alpha GI, Big\rangle \rightarrow \langle \alpha GI\rangle$
12	$\langle \alpha GI, Big\rangle \rightarrow \langle Bi\rangle$	$\langle SU, Big\rangle \rightarrow \langle Big\rangle$
13	$\langle Big, DPP4i\rangle \rightarrow \langle SU, Big, DPP4i\rangle$	$\langle \alpha GI, Big\rangle \rightarrow \langle Big\rangle$
14	$\langle SU, Big, THZ\rangle \rightarrow \langle SU, Big, THZ, DPP4i\rangle$	$\langle \alpha GI, THZ\rangle \rightarrow \langle \alpha GI\rangle$
15	$\langle SU, \alpha GI\rangle \rightarrow \langle SU, \alpha GI, Big\rangle$	$\langle SU, \alpha GI, Big\rangle \rightarrow \langle SU, Big\rangle$
16	$\langle SU, THZ\rangle \rightarrow \langle SU, Big, THZ\rangle$	$\langle SU, \alpha GI, Big\rangle \rightarrow \langle SU, \alpha GI\rangle$
17	$\langle \alpha GI\rangle \rightarrow \langle DPP4i\rangle$	$\langle SU, \alpha GI, THZ\rangle \rightarrow \langle SU, \alpha GI\rangle$
18	$\langle \alpha GI, Big\rangle \rightarrow \langle \alpha GI, Big, DPP4i\rangle$	$\langle SU, \alpha GI\rangle \rightarrow \langle \alpha GI\rangle$
19	$\langle THZ\rangle \rightarrow \langle DPP4i\rangle$	$\langle SU, THZ\rangle \rightarrow \langle THZ\rangle$
20	$\langle SU, Big\rangle \rightarrow \langle SU, \alpha GI, Big\rangle$	$\langle RaIS, \alpha GI, Big\rangle \rightarrow \langle \alpha GI, Big\rangle$

2010, the clinical physician would like to understand the impact of the event onto the physician behavior. Towards these question, the clinical physician has an underlying assumption that the released new drug impact the diminishing usage of SU and physicians behavior. In order to clarify the assumption, we performed experiments to answer the following three questions onto both datasets (before and after 2010):

Question No. 1. Is there any transition(s) from medication that previously contain SU into medication without SU?

Question No. 2. If there were such pattern stated in No. 1, how many of patient having such pattern?

Question No. 3. If there is such pattern stated in No. 1, what is the medicine that replace SU?

We use a pattern model of singleton with flexible distance $\langle X \rightarrow^{+} Y \rangle$ to answer Question No.1 so that we could show medication transitions from using SU to not using SU in the future. The result for Question No. 1 is shown in Table 3.7, which lists top 20 of singleton pattern based on the support value with flexible distance. The results show that there are pattern medication transition from using SU to not using SU in both datasets (before and after 2010 dataset). From pattern before 2010, Biguanide is used in 15 out of 20 top pattern. This result is higher compared to other medication type that are as follows: RaIS, αGI, and THZ are used respectively in 8, 10, and 6 out of 20 top pattern. Additional result is that medication that previously contain Biguanide is maintained in the next medication as shown in row 4, 9, 11-15, and 17-20. Results in pattern after 2010, DPP4i is used in 16 out of 20 top pattern. This number is significantly higher than other medication types, which are as follows: RaIS, α GI, Biguanide, and THZ are used respectively in 5, 4, 9, and 0 out of 20 top pattern. Unlike pattern before 2010, in pattern after 2010 medication containing Biguanide was not always maintained, such as in row 13.

For Question No. 2, we first use a hybrid pattern model to reveal pattern that start with monotherapy SU to any medication not using SU ($X \rightarrow^{+} (\neg X)^{*}$). This pattern is to show the subpopulation of patient that may start the treatment with a monotherapy of SU. The general notion that once the treatment was started with SU, it will be hard to ceased using SU, makes the pattern is interesting to investigate. A conventional pattern model is also used to give a general overview of the patient number with any medication using SU changing to any medication

Table 3.7: Top 20 of singleton pattern with flexible distance with constraints as follows: antecedent must contain medication type 1 (SU) and consequence must be absence from medication type 1.

No	Pattern before 2010	Pattern after 2010
1	$\langle SU \rangle \rightarrow^{+} \langle RaIS \rangle$	$\langle SU \rangle \rightarrow^{+} \langle DPP4i \rangle$
2	$\langle SU \rangle \rightarrow^{+} \langle Big \rangle$	$\langle SU, Big \rangle \rightarrow^{+} \langle Big, DPP4i \rangle$
3	$\langle SU, \alpha GI \rangle \rightarrow^{+} \langle GI \rangle$	$\langle SU, DPP4i \rangle \rightarrow^{+} \langle DPP4i \rangle$
4	$\langle SU, Big \rangle \rightarrow^{+} \langle Big \rangle$	$\langle SU, \alpha GI \rangle \rightarrow^{+} \langle DPP4i \rangle$
5	$\langle SU \rangle \rightarrow^{+} \langle \alpha GI \rangle$	$\langle SU, Big \rangle \rightarrow^{+} \langle Big \rangle$
6	$\langle SU \rangle \rightarrow^{+} \langle RaIS, Big \rangle$	$\langle SU, Big, DPP4i \rangle \rightarrow^{+} \langle Big, DPP4i \rangle$
7	$\langle SU \rangle \rightarrow^{+} \langle \alpha GI, Big \rangle$	$\langle SU \rangle \rightarrow^{+} \langle Big, DPP4i \rangle$
8	$\langle SU \rangle \rightarrow^{+} \langle THZ \rangle$	$\langle SU, Big, DPP4i \rangle \rightarrow^{+} \langle RaIS, Big, DPP4i \rangle$
9	$\langle SU, Big \rangle \rightarrow^{+} \langle RaIS, Big \rangle$	$\langle SU \rangle \rightarrow^{+} \langle Big \rangle$
10	$\langle SU, \alpha GI \rangle \rightarrow^{+} \langle \alpha GI, Big \rangle$	$\langle SU, \alpha GI \rangle \rightarrow^{+} \langle \alpha GI, DPP4i \rangle$
11	$\langle SU, Big \rangle \rightarrow^{+} \langle Big, THZ \rangle$	$\langle SU, THZ \rangle \rightarrow^{+} \langle DPP4i \rangle$
12	$\langle SU, Big \rangle \rightarrow^{+} \langle \alpha GI, Big \rangle$	$\langle SU, DPP4i \rangle \rightarrow^{+} \langle RaIS, DPP4i \rangle$
13	$\langle SU, \alpha GI, Big \rangle \rightarrow^{+} \langle \alpha GI, Big \rangle$	$\langle SU, Big \rangle \rightarrow^{+} \langle DPP4i \rangle$
14	$\langle SU, \alpha GI \rangle \rightarrow^{+} \langle Big, THZ \rangle$	$\langle SU \rangle \rightarrow^{+} \langle RaIS \rangle$
15	$\langle SU, GI, Bi \rangle \rightarrow^{+} \langle < RaIS, \alpha GI, Big \rangle$	$\langle S\alpha GI \rangle \rightarrow^{+} \langle \alpha GI \rangle$
16	$\langle SU \rangle \rightarrow^{+} \langle RaIS, \alpha GI \rangle$	$\langle SU, \alpha GI, Big, DPP4i \rangle \rightarrow^{+} \langle \alpha GI, Big, DPP4i \rangle$
17	$\langle SU, Big, THZ \rangle \rightarrow^{+} \langle Big, THZ \rangle$	$\langle SU, Big \rangle \rightarrow^{+} \langle RaIS, Big, DPP4i \rangle$
18	$\langle SU, Big \rangle \rightarrow^{+} \langle RaIS, Big, THZ \rangle$	$\langle SU \rangle \rightarrow^{+} \langle \alpha GI, DPP4i \rangle$
19	$\langle SU, Big \rangle \rightarrow^{+} \langle RaIS, \alpha GI, Big \rangle$	$\langle SU, Big, THZg \rangle \rightarrow^{+} \langle Big, DPP4i \rangle$
20	$\langle SU, Big \rangle \rightarrow^{+} \langle RaIS, \alpha GI, Big, THZ \rangle$	$\langle SU \rangle \rightarrow^{+} \langle RaIS, DPP4i \rangle$

Table 3.8: Results for Question No 2: Is there any transition from medication that previously contain SU into medication without SU?

Pattern before 2010 (2000-2009)			
$X \rightarrow^{+} Y$	Num	Support	Confidence
$SU \rightarrow^{+} (\neg SU)^{*}$	88	0.04941	0.21256
$SU^{*} \rightarrow^{+} (\neg SU)^{*}$	142	0.07973	0.24232

Pattern after 2010 (2010-2015)			
$X \rightarrow^{+} Y$	Num	Support	Confidence
$SU \rightarrow^{+} (\neg SU)*$	109	0.057428	0.343848
$SU^{*} \rightarrow^{+} (\neg SU)^{*}$	257	0.135405	0.329910

without SU ($X^{*} \rightarrow^{+} (\neg X)^{*}$). Table 3.8 shows the results for Question No 2. The results shows that the number of patient having a monotherapy of SU and then not having SU in the future increased by 25%. However, the confidence is increased significantly by 61%. In addition, the patient number that previously having any medication using SU and then not having SU in the future, is almost doubled from 142 patients to 257 patients and the confidence is increased by 36%. This increase happens only in 5 year after 2010, which means that such medication transition events occurred more frequently after 2010.

And for Question No. 3, the results is shown in Table 3.9 and Table 3.10. With the same reason, we investigate patients previously having a monotherapy of SU and then change to any medication without SU to show the sub populations characteristics. Table 3.9 shows results using for hybrid pattern model and Table 3.10 shows results using conventional pattern model. Comparing the results of before 2010 and after 2010 dataset, similar results are shown on both Table 3.9 and Table 3.10 that is DPP4i is mainly use in any medication combination to cease using SU on after 2010 dataset. For before 2010 dataset, other medications are used fairly, which means that there is no significant differences in support and confidence values. In addition, Biguanide becomes the strong alternative beside DPP4i from Table 3.10. However, for patients previously having monotherapy of SU, DPP4i is the main medication used as shown in 3.9.

Table 3.9: Results for Question No. 3 - Hybrid model: If there is such pattern stated in No. 1, what is the medicine that replace SU?

Pattern before 2010 (2000-2009)			
$X \rightarrow^{+} Y(\neg SU)^{*}$	Num	Support	Confidence
$SU \rightarrow^{+} RaIS(\neg SU)^{*}$	49	0.02751	0.11835
$SU \rightarrow^{+} Big(\neg SU)^{*}$	37	0.02077	0.08937
$SU \rightarrow^{+} \alpha GI(\neg SU)^{*}$	25	0.01404	0.06038
$SU \rightarrow^{+} THZ(SU)^{*}$	17	0.00955	0.04106

Pattern after 2010 (2010-2015)			
$X \rightarrow^{+} Y(\neg SU)^{*}$	Num	Support	Confidence
$SU \rightarrow^{+} DPP4i(\neg SU)^{*}$	92	0.04847	0.29022
$SU \rightarrow^{+} Big(\neg SU)^{*}$	22	0.01159	0.06940
$SU \rightarrow^{+} RaIS(\neg SU)^{*}$	12	0.00632	0.03785
$SU \rightarrow^{+} \alpha GI(SU)^{*}$	9	0.00474	0.02839
$SU \rightarrow^{+} HZ(SU)*$	4	0.00211	0.01261

Table 3.10: Results for Question No. 3 - Conventional model: If there is such pattern stated in No. 1, what is the medicine that replace SU?

Pattern before 2010 (2000-2009)			
$X^{*} \rightarrow^{+} Y(\neg SU)^{*}$	Num	Support	Confidence
$SU^{*} \rightarrow^{+} Big(\neg SU)^{*}$	71	0.03987	0.12116
$SU^{*} \rightarrow^{+} RaIS(\neg SU)^{*}$	66	0.03706	0.11262
$SU^{*} \rightarrow^{+} \alpha GI(\neg SU)^{*}$	60	0.03369	0.10238
$SU^{*} \rightarrow^{+} THZ(\neg SU)^{*}$	37	0.02077	0.06313

Pattern after 2010 (2010-2015)			
$X^{*} \rightarrow^{+} Y(\neg SU)^{*}$	Num	Support	Confidence
$SU^{*} \rightarrow^{+} DPP4i(\neg SU)^{*}$	220	0.11591	0.28241
$SU^{*} \rightarrow^{+} Big(\neg SU)^{*}$	101	0.05321	0.12965
$SU^{*} \rightarrow^{+} \alpha GI(\neg SU)^{*}$	48	0.02529	0.06161
$SU^{*} \rightarrow^{+} RaIS(\neg SU)^{*}$	45	0.02371	0.05776
$SU^{*} \rightarrow^{+} THZ(\neg SU)^{*}$	16	0.00843	0.02053

3.5 Discussion

3.5.1 Pattern Models Interpretation

For the rank similarity analysis, Kendall's Tau takes values between minus one and plus one, with plus 1 meaning the two rankings are identical and -1 meaning one is in reverse of the other [11]. The Kendall's Tau score for FP_{cspade} and $FF_{singleton}$ is 0.42, which means that although FP_{cspade} and $FP_{singleton}$ do have a positive correlation (both rankings have linear correlation), however this correlation is moderately low.

In addition to general characteristic of both result sets, we are able to yield a finding from the medical point of view. Our finding is that by inspecting Table 3.4, the domain expert identify that the transition from Sulfonylurea (SU) to DPP4-inhibitor (DPP4i), as shown in row 7 of FP_{cspade}, is unlikely to happen in high frequency. In the case of the characteristics of Kyoto University Hospital, a physician usually used medication which is the first medication prescribed to the patient as a basic medication. Moreover, when the medication progress, the physician will use the basic medication and combine it with other medication. This behavior is demonstrated by $FP_{singleton}$ pattern in row 11 (⟨SU⟩ → ⟨SU,Big⟩) and row 20 (⟨SU⟩ → ⟨SU,DPP4i⟩) where SU are changed into a dual therapy of ⟨SU⟩ → ⟨SU,*⟩.

The support value of ⟨SU⟩ → ⟨DPP4i⟩ pattern yields a high number in FP_{cspade} set. It is because Apriori algorithm permits partial itemsets and not consecutive order of sequences to support the pattern. Thus, the sequence ⟨*,SU,*⟩* $\rightarrow^{+}$ ⟨*,SU,*⟩ supports ⟨SU⟩ → ⟨DPP4i⟩ pattern. Considering this fact, to make use of FP_{cspade} set, as it is, for a suggestion application based on the support rank value may lead to biased interpretation. For example, in the case of pattern ⟨SU⟩ → ⟨DPP4i⟩ in FP_{cspade}, the user may infer that there are medication transition with high frequency from medication SU to DPP4i, whereas in contrast, it has much lower frequency as recorded in the patient medical history. Compared to $FP_{singleton}$ set, the support value difference of the transition is 0.320293 ($sup_{cspade} - sup_{singleton}$).

Furthermore, based on the result of the first experiment on Table 3.6, we found similar phenomena. This phenomena is demonstrated by $FP_{singleton}$ in row 5 $\langle SU \rangle \rightarrow \langle SU, Big \rangle$. This result is in contrast the conventional method result that pattern $\langle SU, Big\rangle \rightarrow \langle SU\rangle$ is in row 5 of FP_{cspade}, which in $FP_{singleton}$ pattern $\langle SU, Big\rangle \rightarrow \langle SU\rangle$ has a negative correlation (lift < 1) that means the usage of medication pattern $\langle SU, Big\rangle$ does not increase the possibility of the use medication pattern $\langle SU\rangle$.

Having the same result in two case of experiment, we are able to understand that the result set of conventional method should not be interpreted that a certain medication will progress to certain medication combination. However, we could interpret the conventional pattern as a medication combination in the antecedent (the left side) may have a high possibility that in the future it will progress to a medication combination in the consequent (the right side). For example, from Table 3.6 No. 1 ($\langle Big, THZ\rangle^{*} \rightarrow^{+} \langle Big\rangle^{*}$), this conventional pattern may be interpreted that a medication combination of $\langle Big, THZ\rangle$ has a high possibility the next medication will still using $\langle Big\rangle$. This kind of interpretation has a valuable meaning for the clinical physician. However, we are still unable to know for sure whether a medication is stopped from the conventional result set. From the transition $\langle Big, THZ\rangle^{*} \rightarrow^{+} \langle Big\rangle^{*}$, we can not know for sure whether medication "THZ" will be stopped in the future. Such information is also important for clinical physicians and we are able to attain that kind of information from singleton pattern. A singleton pattern, from Table 3.6 No. 1 ($\langle SU, \alpha GI, Big, THZ\rangle \rightarrow^{+} \langle SU, \alpha GI, Big\rangle$), can be interpreted that a medication combination of $\langle SU, \alpha GI, Big, THZ\rangle$ have a high possibility that it will be continued by medication combination $\langle SU, \alpha GI, Big\rangle$. In this pattern, we are able to understand that after medication $\langle SU, \alpha GI, Big, THZ\rangle$, it has high possibility that the medication "THZ" will be stopped. In our proposed method, the time reference is clear (each sequence has a distance by 1 sequence). Hence, the causality is more certain compared to conventional mining. This particular features may benefit the following two possible cases:

Case 1. Identification of adverse drug reaction.

In adverse drug reaction, a medication may cause another disease as reaction of medication combination. In other cases of adverse drug reaction, certain medication may cause different illness. Hence, if the patient's condition progressing in the increasing number of medication combination, then we could capture the adverse drug reaction events.

Case 2. Drug repositioning.

Drug repositioning is an effort to identify whether a certain drug can cure other condition, which is different from its original purpose, by using brute force method. In this case, if the patient condition is diminishing with an addition of certain medication, it is possible that the condition is cured by the added medication.

3.5.2 Principal Finding

From the confirmatory experiments, we are able to reveal the way of physician ceased to use Sulfonylurea (SU). In addition, the released of the drug of DPP4i has an impact on the medication strategy of the physician in selecting the medication to replace SU.

Results from Table 3.7 answer the Question No. 1. The result show that there are frequent patterns of transition from medication using SU to medication without SU. Despite RaIS is the medication type ranked first as a replacement of SU in before 2010 dataset, Biguanide is frequently used in combination with other medication. In addition, RaIS is ranked 14th in after 2010 dataset. In after 2010 dataset, the first rank and the most used medication in any combination is singleton DPP4i. As for Question No. 2, it is answered with the results shown in Table 3.8. The results demonstrate the trend of diminishing usage of SU that the number of medication transition from any medication containing SU to any medication without SU increased significantly. Moreover in a close up, the frequency of monotherapy SU replaced by any medication without SU also increased after 2010.

Furthermore, based on Question No. 3 answers in Table 3.9 and Table3.10, the following insights are attained: Before 2010, there are patterns that replacing SU with other medication types. However the number is fair between those other medication types. Even though, Biguanide and RaIS have high support value, but the numbers only slightly different with other medication type used as a replacement for SU. This is in contrast with way of physician diminishing the use of SU in after 2010. The difference is that DPP4i is most preferable to be used in replacing SU. In addition, the number is significantly different compared to other medication type usage. Other medication type that becomes a strong alternative to replace SU in any medication combination is Biguanide. These results are fit the prior impression from the domain expert. However, in the subpopulation that previously having monotherapy of SU, DPP4i dominate the usage of medication for in any medication combination in replacing SU. Another highlight of the results is based on Table 3.9, which is, there are many cases of medication transition that is from SU to any medication using RAIS. In addition, any medication containing RAIS replacing SU is higher compared to Biguanide in before 2010 dataset. This result is unexpected by the clinical physician. However, it is understandable because RAIS function is similar with SU that is to control the secretion of insulin in the metabolism.

Based on the results, the domain expert is able to confirm the impression of the new released medicines impacts towards the physician behavior to diminish the usage of SU. The impacts are that the way of replacing SU is changing and it is highly dominated by the new released drug (DPP4i). This method can be applied in a wider area of clinical situation, for example, to examine a medication strategy that start with certain medication.

3.6 Conclusion

We proposed new notion of full itemset named singleton into frequent sequential pattern mining method. By incorporating the conventional mining features (flexible distance and subset itemset) into the singleton mining method, the incorporation enables us to obtain fine-grained patterns, which is useful to show a more clear view of the medication strategy in certain subpopulation. Our method enables many physician to understand the changing of using drugs in many area, when launching new drugs. This phenomena is difficult for physician to analyze because of the nature of the long term medication history dataset. Clinical physicians should be benefited with this method.

CHAPTER 4

Medication Strategy Visualization

4.1 Motivation

Effective analysis of time-oriented multivariate clinical data, with the objective of investigating processes and predicting their course, as is important in the case of diabetes, requires the combined use of multiple approaches, including mining the longitudinal clinical data to automatically discover within it meaningful patterns, and enabling the analyst to explore the result [28]. Conventional methods presents result sets of mining activity in tabular manner based on ranking functions. However, this method may discourage an analyst to explore the result set because the frequent patterns may be in a great number. As found in [27], when given the results in the a tabular manner, an analyst will only give a high attention on the first hundred of results and the bottom results. Visual analytic (VA) methods attempt to bridge this requirement of analyzing the result in effectively manner.

The first attempt of visual exploration systems in medical domains focused mostly on the visualization of raw longitudinal data for individual or multiple patient records. Common goal of previous studies are development of innovative interfaces, graphical metaphors, and exploration capabilities, rather than on the discovery of actual new knowledge [2]. Recent visual exploration systems include additional capabilities for sophisticated interactive exploration of multiple patients data for cohort studies. However, most VA systems focus on raw data, such as a time series of laboratory test results, rather than on its interpretations. They include neither an underlying domain-specific knowledge base that formally represents the explored concepts and the relationships among them.

Our study similar is with [28] that is to analyze longitudinal datasets. [28] developed Visual Temporal Analysis Laboratory (ViTA-Lab). It is a framework that combines data-driven temporal data mining techniques, with interactive, query-driven, visual analytical capabilities, to support, in an integrated fashion, an iterative investigation of time oriented clinical data and of patterns discovered in them. However, the focus of their study is temporal relationship patterns as described by [3]. This is different with our study, which emphasize on medication strategy marked by medication transition events listed in Section 2.3.3. A medication strategy represents not only systematic actions by the physician but also the reasoning behind them that includes the prior condition and the objective of the actions. Our aim in VA is to provide visualization that enables physician to examine the medication strategy from the mining activity result set.

We focus our study in the usage of directed graph to develop visualization that represent the medication strategy. Graph has been used to visualize varying datasets, ranging from social network[48], web structure[9], traffic and molecular communication [24]. Graph is a collection of nodes (vertices) and edges (i.e., links that connect the nodes) [15]. Nodes and edges are used to visualize the data. In our case, the nodes are the medication combination and edges are the medication transition events. In this chapter, two type of graph visualizations is discussed.

4.2 Related Works

4.2.1 Medication Strategy in Type 2 Diabetes

Type 2 Diabetes is a common chronic disease with the highest prevalence number based on WHO documentation. In the case of diabetes, the selection of pharmacotherapy is considered essential [51]. The appropriate combination of medications should be selected in accordance with the patient conditions.

Our study consider medication transition events are essential because they are marker when the patients condition changes and the physician needs to modify the treatment. As listed in Chapter 2, medication transition events may be in the form of adding, stopping, switching, and continuing medications. In addition, the transition can also be in the form of decreasing or increasing the dosage. As for the patient's condition, there are several patients indicator used by physician to decide which strategy use. Two indicators discussed in this study is A1c value and eGFR.

The A1c measures an average blood glucose for the past 2 to 3 months. Diabetes is diagnosed at an A1C of greater than or equal to 6.5. GFR is Glomerular Filtration Rate and it is a key indicator of renal function of the kidney. At the onset of diabetes, the kidney grows large and the GFR becomes disturbed. eGFR is estimated GFR and is a mathematically derived entity based on a patients serum creatinine level, age, sex and race.

4.2.2 Visualization using Graph

Definitions used in a graph theory is given as the following [13]: A graph is a set V together with a relation on V. A graph $G = (V, E)$ is a pair of sets, V is a set of nodes (vertices), E is a set of edges (arcs or links). An edge $e(u, v)$, with, $e \in E$ and u,$v \in V$, is a pair of vertices. If the relation on V induced by E is symmetric; we call such a graph undirected. If the pair of vertices in an edge is ordered, G is denotes as directed graph or digraph. Direction is denoted by saying, with respect to a node, that an edge is incoming or outgoing. A graph is weighted if each of its edges is associated with a real number. An unweighted graph is equivalent to a weighted graph whose edges all have a weight of 1. A graph is complete if there exists an edge for every pair of vertices. If it has n vertices, then a complete graph has $n(n-1)/2$ edges. A loop is an edge with $u = v$. A path is a list of successively adjacent, distinct edges. Let $< e_1, ..., e_k >$ be a sequence of edges in a graph. This sequence is called a path if there are vertices $< v_1, ..., v_k >$ such that $e_1 = (v_{i-1}, v_i)$ for $i = 2, ..., k$. A path is cyclic if a node appears more than once in its corresponding list of edges. A graph is cyclic if any path in the graph is cyclic and acyclic if there are no cyclic path in the graph. A tree is a graph in which any two nodes are connected by exactly one path. Trees are thus acyclic connected graphs. Trees may be directed or undirected. A tree with one node labeled root is a rooted tree.

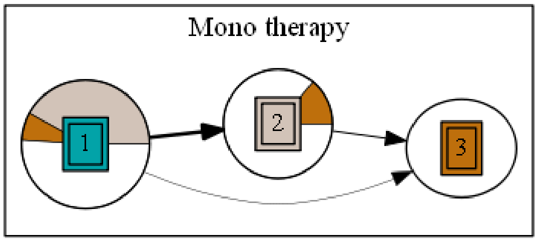


Figure 4.1: A sample of pattern representation.

4.3 Methodology

4.3.1 Medication Theraphy Transition Graph

As mentioned in Section 4.1 (Motivation), one of our aims is to learn about the underlying reasoning behind the physician behavior in changing the medication. We are interested in the top n with the highest support patterns. The mining result is shown in the form of directed graph (using GaphViz 2.38.0). Figure 4.1 displays a sample of pattern representation. The nodes represent the singleton patterns. Circle nodes (node 1 and node 2) show that treatment patterns is among top n patterns with the highest support. The larger the circle means the higher the support value. Oval node (node 3) means that the treatment pattern is not among the top n patterns. The arrows represent the 1-sequence patterns. The thicker the arrows the higher the support value of the sequence pattern.

In order to learn the underlying reason of the physician, patient clinical indicators are used, that is the A1c and creatinin serum results. We transformed the creatinine serum value into the eGFR value, which the physician uses to understand the patient's renal function condition [23]. The clinical indicator values used are lab results indicated in Figure 4.2. In addition, we abstracted the A1c lab test results into three categories: *ltr*: lower than the range (< 7), *ir*: within the range ($\geq 7 \wedge \leq 8$), and *gtr*: greater than the range (> 8). We abstracted the eGFR into two categories as follows: *ltr*: lower than the level (≤ 60) and *ir*: within range (> 60). The coloring assignment of pair variation beetwen A1c and eGFR vales is shown in Table 4.1.

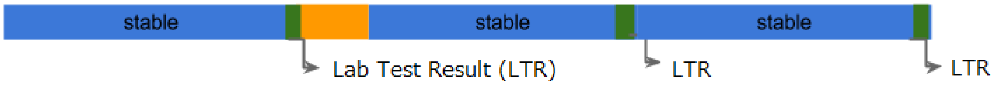


Figure 4.2: Lab test Result used in the data mining.

Table 4.1: The coloring assignmentof pair variation between A1c and eGFR.

⟨A1c,eGFR⟩	Color	Annotation
⟨ltr,ir⟩	green	A1c $<$ 7 and eGFR $>$ 60
⟨gtr,ir⟩	blue	A1c $>$ 8 and eGFR $>$ 60
⟨ir,ir⟩	orange	$7 \leq$ A1c ≤ 8 and eGFR $>$ 60
⟨ltr,ltr⟩	purple	A1c $<$ 8 and eGFR $\leq$ 60
⟨gtr,ltr⟩	red	A1c $>$ 8 and eGFR $\leq$ 60
⟨ir,ltr⟩	purple	$7 \leq$ A1c ≤ 8 and eGFR $\leq$ 60

4.3.2 Medication Trajectory Graph

From a longitudinal dataset, medication strategy can also show medication pathways. This information is important for clinical physicians to understand not only long term strategy but also the patient condition's pathways. We are interested in developing a medication trajectory graph from 1-sequence patterns produced by singleton mining.

The requirements of the medication trajectory graph are as the following:

1. The graph should be an acyclic-rooted tree graph with left to right direction.
2. Nodes and edges are frequent patterns in the form of singleton and 1-sequence pattern produced by singleton mining
3. A node is a singleton, which represents combination of medication and an edge is a sequence of adjacent singletons, which represents a medication transition event. The node and edge are associated with the support of the pattern.
4. The root should be a monotherapy and then, propagate into dual therapy, triple therapy and so on.

An example of medication trajectory model is shown in Figure 4.3. Figure 4.3 shows a three levels of tree graph. The first level is the root (monotherapy), the second level should be dual therapies, and the third level should be triple therapies. And as demonstrated by Figure 4.3, there are no loops in the graph (i.e., incoming edge from node in the later level).

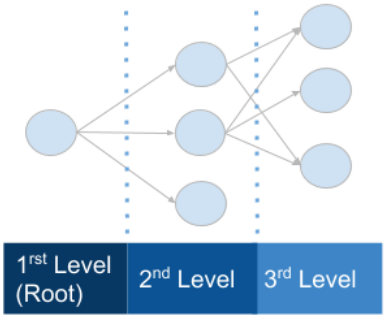


Figure 4.3: An example of medication trajectory model.

To enable dynamic visualization, we develop the graph using PHP, which is a general-purpose scripting language [17], and Vis.js, which is a javascript library for dynamic, browser based visualization [18]. The algorithm is shown in Algorithm 1. The weight retrieval is not described in the algorithm. However, it can be easily retrieve along with the singleton and 1-sequence patterns retrieval respectively are pushed into N and E hash respectively to the nodes and edges.

4.4 Experiments and Discussion

In this section, two experiments are conducted to show each visualization strong points. The first experiment is to display Top-k medication transition pattern and the second experiment is to display two medication trajectory of two period of time partitioned from the main dataset.

4.4.1 Displaying Top-K Medication Transition Pattern

We analyzed Type 2 diabetes patient EMR provided by Kyoto University Hospital. The medical history spans September 2000 - April 2015. We are interested in finding the medication pattern with the highest support value. We exclude the injection medicine (insulin and GLP1-Receptor Agonist). After the exclusion, we have a raw data set that covers 6,573 patients with a total of 224,269 records.

Our analysis method is displayed in Figure 4.4. The mining process consists of three sub-procedures (medical episode reconstruction, lab test result abstraction,

Algorithm 1 Build Medication Trajectory Graph (MTC)

```
    Input:   a set of singletons S and 1-sequence patterns T
    Output:   trajectory model
procedure Build MTC
Select a monotherapy root from S and maximum combination number maxLevel from T
Initialize previous level node prevAnt = root
Initialize nodes N and edges E as hashes
for $level = 2; level \leq maxLevel; level++$ do
    select 1-sequence pattern(s) e, destination node(s) con(s) from T having source node(s) ant(s) equal to prevAnt(s) and con(s) is not the same with the prevAnt(s) from T
    if con is not in N then
        push con into N
    end if
push e into E
end for
build directed graph with N, E data
endProcedure
```

and sequential mining). Using the medication reconstruction to identify the stable period, we are able to compress the search space into 53,444 records (23.83% compared to the raw data).

The results of our analysis are presented in Figure 4.5 (by using an ϵ value of 14 days and δ value of 90 days). One highlight of the results shown in the enlargement of monotherapy-cluster of Figure 4.5. There are four parts of the pie chart with the color of Node 6 (DPP4-i) in Node 1(SU). It means that there are four edges that come out from Node 1 (SU) to Node 6 (DPP4-i) (1-sequence pattern of $\langle$SU$\rangle \rightarrow \langle$DPP4-i$\rangle$). There are 2 purple edges (7 patients with combination $\langle$ltr,ltr$\rangle$ and 12 patients with combination $\langle$ir,ltr$\rangle$), one orange edge (8 patients with combination $\langle$ir,gtr$\rangle$) and one green edge (8 patients with combination $\langle$ltr,gtr$\rangle$).

As shown in Figure 4.5, showing the mining results as they are can be overwhelming because many of the transitions are actually acceptable. Therefore, we added a transition filter, as shown in Figure 4.4. One of the filter condition

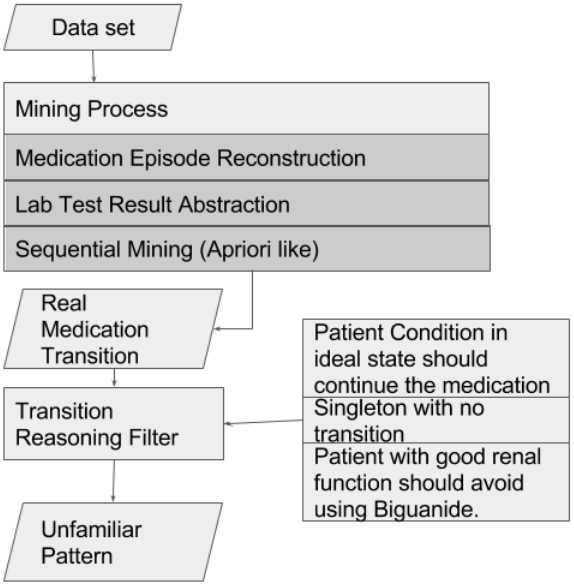

Data set
Mining Process
Medication Episode Reconstruction
Lab Test Result Abstraction
Sequential Mining (Apriori like)
Real Medication Transition
Patient Condition in ideal state should continue the medication
Singleton with no transition
Patient with good renal function should avoid using Biguanide.
Transition Reasoning Filter
Unfamiliar Pattern

Figure 4.4: Analysis method.

is that patient condition in ideal state should continue the medication. Using this condition, a clinical physician can observer some green edges. For example, one green edge comes out from singleton 1 (SU) towards singleton 6 (DPP4i), as indicated by the red arrow in Figure 4.5. This green edge means that therapy transitions occurred, though both the patient conditions were ideal conditions (A1c < 7) and the renal functions were good (eGFR > 60). This green edge is an unfamiliar transition because based on the medical guidelines, it is recommended that the medication be continued if the target control is achieved. In [51], the recommended target control is less than 7.0%. This event insinuates that the medication transition may not only be caused by the patient condition and that one possible driver of this decision could be the newly released DPP4-i medicine.

4.4.2 Comparing Two Periods of the dataset using Medication Trajectory Graph

As explained in section 4.3.2, clinical physicians are also interested in understanding the medication pathways that is medication transitions from monotherapy to multitherapy. This visualization hold a valuable information about the long term strategy as the patient condition progressing. We use stable period sequences of Type 2 diabetes patients, which are identified from medication episodes construc-

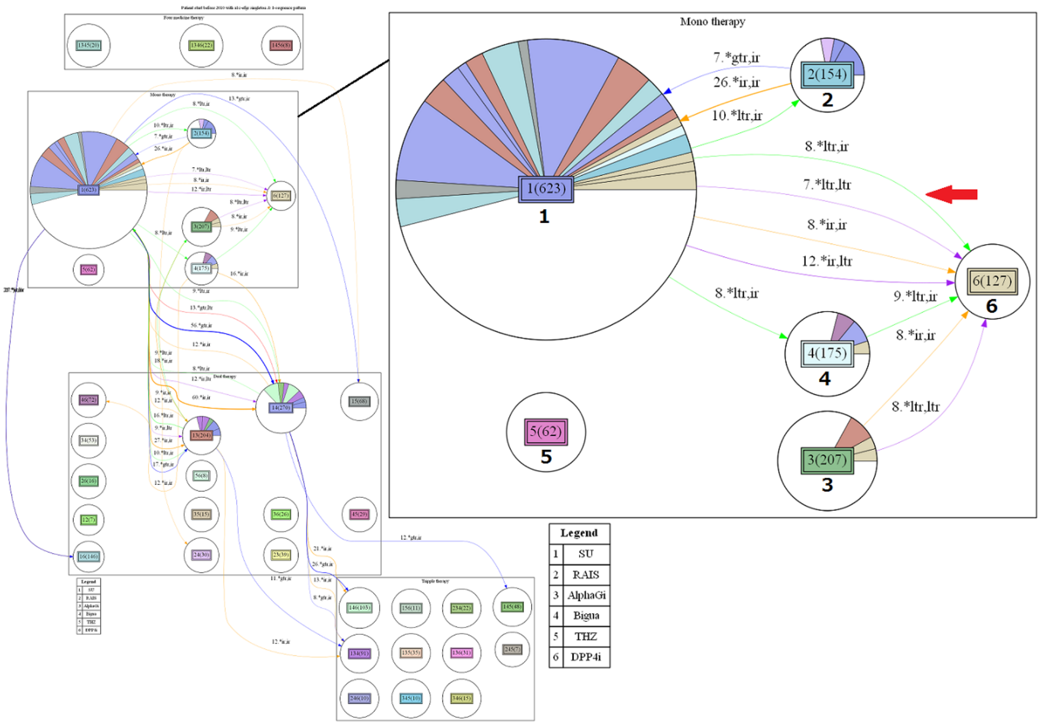


Figure 4.5: Medication therapy transition graph of top 75 support value for patient start before 2010.

tion framework on medical history provided by Kyoto University Hospital along with the approval from the Ethics Review Board of The Medical School of Kyoto University. Figure 4.6 shows the framework for constructing the medication trajectory graph.

In this experiment, we conduct two visualizations by partitioning the dataset into two parts: dataset prior 2010 and after 2010. The reason for partitioning on that particular year is because there was a release of new diabetes medicine (DPP4i medication type number 6) in 2010. We would like to investigate the physician strategy before and after the release of DPP4i. Prior 2010, we have stable period sequences from 1781 patients, while after 2010, we have 1898 patients. In addition, the nodes and edges are patterns having minimum support of 0.001.

The results of the visualisation prior 2010 and after 2010 are shown by Figure 4.7 and Figure 4.8, respectively. As described in section 4.3.2, the root is the monotherapy, the second level is the dual therapies, the third level is the tripple therapies, and the fourth level is therapies with 4 medications. The label inside

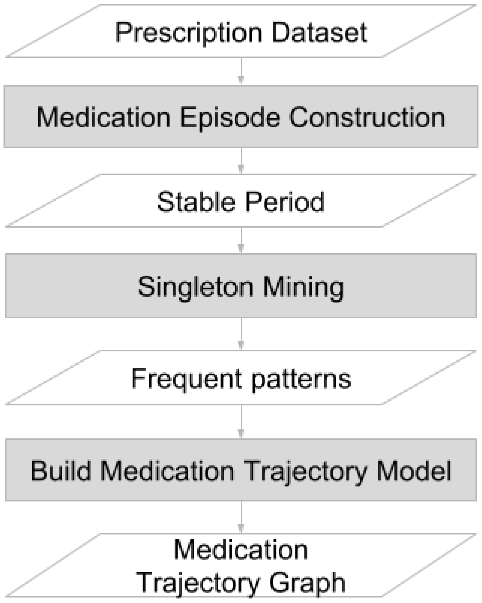

Prescription Dataset
Medication Episode Construction
Stable Period
Singleton Mining
Frequent patterns
Build Medication Trajectory Model
Medication Trajectory Graph

Figure 4.6: Framework for constructing the medication trajectory graph for the experiment.

the nodes shows the medication combination and the number of patient having the pattern.

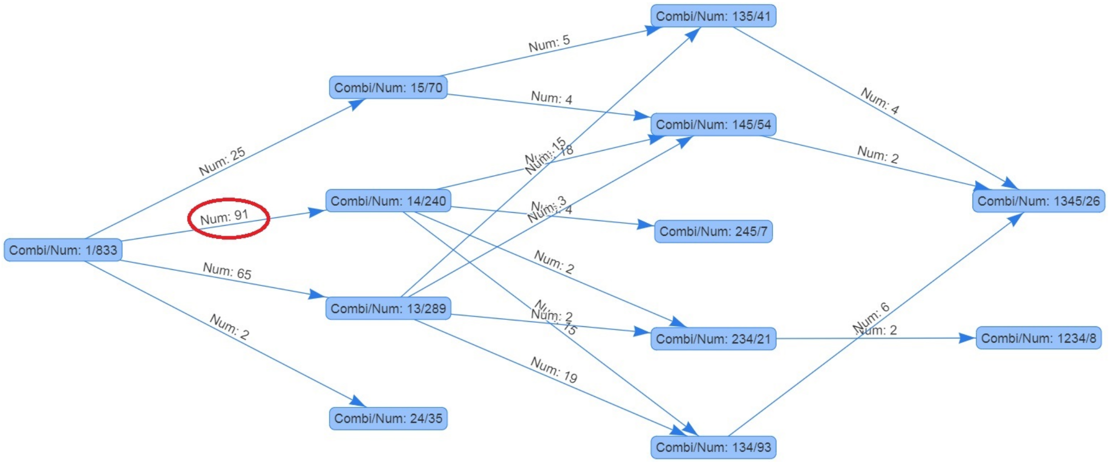


Figure 4.7: Medication trajectory graph for prior 2010.

From Figure 4.7 in root node, the medication combination is medication type 1 that is Sulfonylurea (SU) and there are 833 patients prescribed with SU. The edge from 1 to 14 is a transition event from medication type 1 (SU) to medication

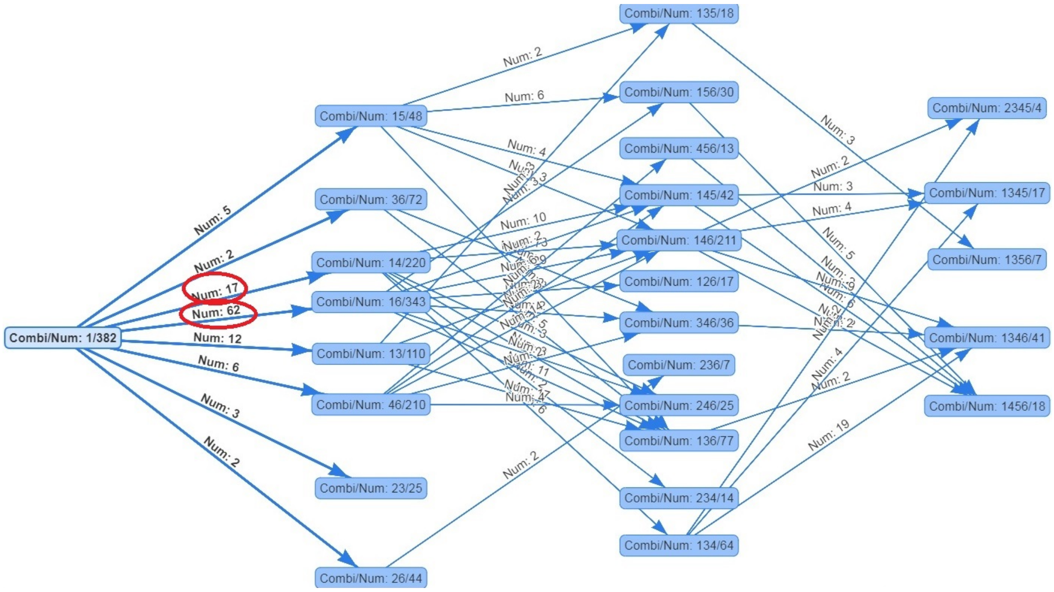


Figure 4.8: Medication trajectory graph for after 2010.

combination of type 1 (SU) and 4 (Biguanides). This transition pattern is occured in 91 patients.

Visualizations of patterns from dataset prior 2010 and dataset after 2010 show that there is a significant difference in number of nodes in each level. In retrospect of a medical doctor from Kyoto University Hospital, the condition shown by Figure 4.8 reflects that physicians strategy was activated by the new released medicine. Medication combinations used by physicians after 2010 vary more compared to prior 2010. Medication type no. 6 is greatly used in new combination medications in each level. Furthermore, comparing the transition from the root node to the second level, node 14 has the highest number of occurrences in Figure 4.7 (circled by red color). In contrast, that number is decreased sharply in Figure 4.8. This situation resembles the physicians choice of medication at the time prior and after the new medicine released that is prior 2010, physicians who started with medication type 1 (SU), when changed the medication into dual therapy was likely to use combination of medication type 1 (SU) and type 4 (DPP4-i) as shown in Figure 4.7. As for after 2010, the medication combination that is likely to use by the physician is combination of medication type 1 (SU) and 6 (DPP4i). The transition from SU to SU and medication combination of Type 1(SU) and 6 (DPP4-i) becomes the highest as shown in Figure 4.8 (circled by red color).

Based on the result, the clinical physician is able to understand the physicians activities before and after the new drug was released. In addition, the advantage of this method is that we are able to show a medication trajectory from monotherapy to multitherapy, which shows the physician preference in medication selection as the patient condition progress.

4.5 Conclusion

We proposed visualization methods to present the result set that is frequent medication strategy patterns. Using medication transition graph, patient's clinical condition is used to understand some extent of physician reasoning. Medication trajectory graph shows strategic pathway taken physicians when patient condition progressing. Our results show that using graph visualization made the clinical physician easier to examine the result set of medication strategy compared to the tabular presentation. This is different with market basket case that mainly used ranking list in tabular manner for domain analyst to examine the result set.

CHAPTER 5

Conclusions and Future Directions

In this thesis, we presented a practical way to conduct a retrospective analyses on long-term medication history in the form of prescription records. This retrospective longitudinal studies have been attracting a high interest from clinical physicians, especially in epidemiology communities. Many benefits are offered by these type of studies, for example to learn the relationship between risk factors, the development of diseases and the outcomes of treatments over different period of time. In addition, the increasing volume of medication history accumulated by the health care provider has provided an new opportunity to conduct retrospective longitudinal analyses using data driven tools. This data driven study on clinical setting has unique issues compared to common market basket cases. Currently, we have addressed several issues on medication strategy analyses from a long-term medication history. The main contributions are summarized as follow:

1. We presented data construction framework from long-term multitherapy prescription records for retrospective database analysis for observing medication transition events. This framework adapts the notions of time error margin ϵ to assign a more flexible Allen's temporal relation [3] between neighbouring prescriptions. Based on the temporal relation assignment, the framework treats adjacent prescriptions according to the construction

rules that we developed with regarding treatment characteristics of the disease and patient's behavior in clinical setting. We define a longer form of reconstructed prescriptions as a medication episode and a stable period as a medication episode having longer duration than a period of time for a physicians to see whether a medication is effective or not. We showed that our framework named medication episode construction framework allows to reducing repetitive medication episode while preserving prescriptions information in a multitherapy dataset. This is important for the longitudinal analysis of chronic diseases, particularly to observe the strategic actions by physicians to achieve ideal condition for the patients.

2. We presented a novel notion of a frequent pattern model that is a singleton pattern. A singleton pattern is defined as a pattern of a full itemset, which is an itemset that is contained equally in at least one itemset of the sequence data. In addition, we define singleton pattern mining task to extract full adjacent itemsets. This method eases the causality analyses between itemsets in the frequent sequence pattern result set compared to the conventional pattern mining method [1] that features partial/subset itemset and non-consecutive sequence are able to support the frequent pattern candidate. Furthermore, by incorporating the conventional mining features (subset itemset and flexible distance) into the singleton mining method, we demonstrated that finer-grained patterns are obtained, which is useful to allow a clinical physician answers deeper research questions.

3. We presented directed graph based visualizations named medication therapy transition graph and meedication trajectory graph. The first graph provides information of existing transition between medication therapy inside the dataset and patient condition prior the transition. Using this information, a clinical physician is able to infer physicians reasoning in adjusting the medication and medication strategy as the disease progress. The second graph provide information of medication strategy as the diseases progress. This graph provide valuable information when used as a mean to compare medication strategy between different periods of time. Information presented in our visualization differs with [32] that focuses to provide temporal relation information of the observed parameter.

Current proposed methods are applicable in a more general chronic condition where clinical physician needs to answer research question using long-term medication history. The study onto long-term medication history is essential to start comprehensive researches for the improvement and development of health care. There are however some limitations of our construction framework and pattern mining techniques, which our proposed methods inherit.

1. Information preserved by the construction framework are as stated by the prescription records. In our case, we only use a dataset from a single hospital. And a patient may go to other clinic because one and other reasons. This condition could produce blank periods that we defined as a period of time when we do not have the information. And with the existence of the blank period, it may required additional consideration in further analyses, such as stated in [46].

2. As other frequent pattern methods, our method suffers the same disadvantage that is the possibility of producing a great number of frequent patterns. This condition may happen when observed items are large. However, in real clinical condition, the number of medicine related to a health condition may not varied much. Especially, with the publication of medical guidelines, recommended treatments are limited in number. Therefore, some modification may be needed when applied in other application, such as item abstraction for frequently found together items as proposed by [29].

We now outline some related open questions and research opportunities.

1. **Analyses with granularity.** The current medication episode framework is able to preserve information from the prescription records, such as medication type, medication name, duration and dosage. However, the analyses' example conducted in this thesis are still limited to the usage of medication types. Extending the data usage in the analyses will provide analyses with levels of detail that is needed by clinical physicians.

2. **Frequent pattern mining with precission.** As we have shown in Chapter 3, that with its features, conventional mining method provides a general view. Applications, such as identification of adverse drug rection and drug repositioning, require a higher degree of precission. Singleton mining extracts full itemsets out of the dataset. This feature shows a potential to be used in these application to identify patient condition changes, such as an

additional/progressed health condition when certain medication is added in adverse drug and a decreased/stopped health condition in applying certain medication in drug repositioning. The deployment of our proposed method in these application will also provide opportunity to answer an important question that is to investigate the implementation of our proposed method in larger observed parameters.

3. **Knowledge based visual analytics.** The current visual analytic in Chapter 4 provides means for analysts to effectively explore the result set. It will provide to enhancement of the analyses ability by allowing usage of the obtained knowledge by the analyst from the current visualization as a query to analyze other datasets. Deployment of intelligent visual based analyses promises a good research future direction.

4. **Data interoperability.** With the current development in personal health application, it is not only health care provider that records health condition and medication intake. Benefit can be harness from these applications by ensuring data interoperability between health care provider's data center and personal application. Data from personal application can be used to complement unknown information as presented in the case of blank period. The same problem may arise in effort to collect dataset from multi health care provider. Other issues needed to be dealt in this case are data privacy and security.

Acknowledgements

In this section, I would like to express my sincere gratitude to my supervisor Prof. Masatoshi Yoshikawa and co-supervisors Prof. Tomohiro Kuroda and Dr. Akihiro Hamasaki, for their constant support, guidance and motivation. It would never have been possible for me to take this work to completion without their incredible support and encouragement. I benefited greatly from many fruitful discussions with them.

I value the discussion with Associate Professor Kazuya Okamoto and Lecturer Dr. Osamu Sugiyama in our research meeting. I have often look forwarded for their valuable comments.

I would like to thank the faculty member of Yoshikawa and Ma Laboratory especially Associate Professors Qiang Ma, Yasuhito Asano, and Adam Jatowt, Lecturer Dr Makoto P. Kato, and Assistant Professors Toshiyuki Shimizu and Yang Cao for their constructive questions and helpful discussions on my effort to conduct my research. I would also like to thank Yuichi Sugiyama, who is my predecessor in this study of long-term prescription records and Ryousuke Sawano for his help in extracting the anonymized dataset. I thanks sincerely for Yiling Dai, Sora Lim and Hideaki Ohashi for the valuable support and discussion. I want to say thank you for our laboratory secretaries Rika Ikebe and Yoko Nakahara for their sincere help.

I can not forget the valuable conversation and discussion of Prof. Rakesh Agrawal that provide me with wider perspective of possible future research directions in his class.

I am truly grateful to my parents for their immence love and support.They have always encouraged me to explore my potential and pursue my dreams. They helped me a lot to reach this stage in my life. I would like to thank my brothers and sisters for supporting and keeping me motivated throughout this work. Special thank you is dedicated to my husband Jun-Bapake, which without his full support and loving encouragement, I would not be able to embark this study. Sweet thank you is also for my two sons Alkhaz and Akhtar for their gift of laughter.

I wish to thank many other people whose names are not mentioned here but this does not mean that I have forgotten their help and support.

In addition, I would like to acknowledge Indonesian Endowment Fund for Education (LPDP) for providing a full scholarship and Indonesian Institue of Sciences (LIPI) for providing a recommendation and permission to my study. This work was also supported by JSPS KAKENHI Grant Number JP15H02705.

And finally, I praise "*Alhamdulillahi rabbil 'alamin*" to Allah Subhanahu Wa Ta'ala that has been my source of energy and spirit through out this work. With His permission and *rahmah*, I am able to complete this work.

Purnomo Husnul Khotimah, August 2018

References

[1] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 3–14. IEEE, 1995.

[2] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

[3] James F Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[4] Hadi Banaee and Amy Loutfi. Data-driven rule mining and representation of temporal patterns in physiological sensor data. *IEEE journal of biomedical and health informatics*, 19(5):1557–1566, 2015.

[5] Iyad Batal, Dmitriy Fradkin, James Harrison, Fabian Moerchen, and Milos Hauskrecht. Mining recent temporal patterns for event detection in multivariate time series data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 280–288. ACM, 2012.

[6] Michele Berlingerio, Francesco Bonchi, Fosca Giannotti, and Franco Turini. Mining clinical data with a temporal dimension: a case study. In *Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on*, pages 429–436. IEEE, 2007.

[7] Stephanie Bernell and Steven W Howard. Use your words carefully: what is a chronic disease? *Frontiers in public health*, 4:159, 2016.

[8] Isabelle Bichindaritz and Stefania Montani. *Case-Based Reasoning: 18th International Conference, ICCBR 2010, Alessandria, Italy, July 19-22, 2010 Proceedings*, volume 6176. Springer, 2010.

[9] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer networks*, 33(1-6):309–320, 2000.

[10] Hahsler M Buchta C and Buchta MC. *Package arulesSequences*, 2016.

[11] Ben Carterette. On rank correlation and the distance between rankings. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 436–443. ACM, 2009.

[12] Edward Joseph Caruana, Marius Roman, Jules Hernández-Sánchez, and Piergiorgio Solli. Longitudinal studies. *Journal of thoracic disease*, 7(11):E537, 2015.

[13] Chun-houh Chen, Wolfgang Karl Härdle, and Antony Unwin. *Handbook of data visualization*. Springer Science & Business Media, 2007.

[14] S Concaro, L Sacchi, C Cerra, P Fratino, and Riccardo Bellazzi. Mining health care administrative data with temporal association rules on hybrid events. *Methods of information in medicine*, 50(02):166–179, 2011.

[15] Diane J Cook, Lawrence B Holder, and Nikhil Ketkar. Unsupervised and supervised pattern learning in graph data. *Mining Graph Data*, pages 159–180, 2006.

[16] Felicia Cosman. Long-term treatment strategies for postmenopausal osteoporosis. *Current opinion in rheumatology*, 30(4):420–426, 2018.

[17] Wolfgang Gaebel, Hans-Jürgen Möller, Gerd Buchkremer, Christian Ohmann, Mathias Riesbeck, Wolfgang Wölwer, Martina Von Wilmsdorff, Ronald Bottlender, and Stefan Klingberg. Pharmacological long-term treatment strategies in first episode schizophrenia. *European archives of psychiatry and clinical neuroscience*, 254(2):129–140, 2004.

[18] Helga Gardarsdottir, Patrick C Souverein, Toine CG Egberts, and Eibert R Heerdink. Construction of drug treatment episodes from drug-dispensing

histories is influenced by the gap length. *Journal of clinical epidemiology*, 63(4):422–427, 2010.

[19] Jesper Hallas and Henrik Støvring. Templates for analysis of individual-level prescription data. *Basic & clinical pharmacology & toxicology*, 98(3):260–265, 2006.

[20] Frank Höppner. Learning temporal rules from state sequences. In *IJCAI Workshop on Learning from Temporal and Spatial Data*, volume 25, 2001.

[21] Jingshan Huang, Jun Huan, Alexander Tropsha, Jiangbo Dang, He Zhang, and Min Xiong. Semantics-driven frequent data pattern mining on electronic health records for effective adverse drug event monitoring. In *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*, pages 608–611. IEEE, 2013.

[22] Weiyi Huang, RL Castelino, and GM Peterson. Metformin usage in type 2 diabetes mellitus: are safety guidelines adhered to? *Internal medicine journal*, 44(3):266–272, 2014.

[23] Enyu Imai, Yoshinari Yasuda, and Hirofumi Makino. Japan association of chronic kidney disease initiatives (j-ckdi). *Japan Med Assoc J*, 54:403–405, 2011.

[24] Satoru Iwasaki and Tadashi Nakano. Graph-based modeling of mobile molecular communication systems. *IEEE Communications Letters*, 22(2):376–379, 2018.

[25] Po-shan Kam and Ada Wai-Chee Fu. Discovering temporal patterns for interval-based events. In *International Conference on Data Warehousing and Knowledge discovery*, pages 317–326. Springer, 2000.

[26] Purnomo Husnul Khotimah, Yuichi Sugiyama, Masatoshi Yoshikawa, Akihiro Hamasaki, Kazuya Okamoto, and Tomohiro Kuroda. Revealing oral medication patterns from reconstructed long-term medication history of type 2 diabetes. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 5599–5603. IEEE, 2016.

[27] Martin Kirchgessner, Vincent Leroy, Sihem Amer-Yahia, and Shashwat Mishra. Testing interestingness measures in practice: A large-scale analysis of buying patterns. *arXiv preprint arXiv:1603.04792*, 2016.

[28] Denis Klimov, Alexander Shknevsky, and Yuval Shahar. Exploration of patterns predicting renal damage in patients with diabetes type ii using a visual temporal analysis laboratory. *Journal of the American Medical Informatics Association*, 22(2):275–289, 2014.

[29] Marion Leleu, Christophe Rigotti, Jean-François Boulicaut, and Guillaume Euvrard. Go-spade: mining sequential patterns over datasets with consecutive repetitions. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 293–306. Springer, 2003.

[30] Nancy P Lin, Hung-Jen Chen, and Wei-Hua Hao. Mining negative sequential patterns. In *Proc. of the 6th WSEAS International Conference on Applied Computer Science, Hangzhou, China*, pages 654–658, 2007.

[31] Fabian Mörchen. A better tool than allens relations for expressing temporal knowledge in interval data. In *Workshop on Temporal Data Mining at the Twelveth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 25–34, 2006.

[32] Robert Moskovitch and Yuval Shahar. Medical temporal-knowledge discovery via temporal abstraction. In *AMIA*, 2009.

[33] Vishal S Motegaonkar and Madhav V Vaidya. A survey on sequential pattern mining algorithms. *International Journal of Computer Science and Information Technologies*, 5(2):2486–2492, 2014.

[34] Mitsuyoshi Namba, Toshio Iwakura, Rimei Nishimura, Kohei Akazawa, Munehide Matsuhisa, Yoshihito Atsumi, Jo Satoh, Toshimasa Yamauchi, et al. The current status of treatment-related severe hypoglycemia in japanese patients with diabetes mellitus: A report from the committee on a survey of severe hypoglycemia in the japan diabetes society. *Diabetology International*, 9(2):84–99, 2018.

[35] R.B. Nelsen. Kendall tau metric, encyclopedia of mathematics, 2011. http://www.encyclopediaofmath.org.

[36] Lars Hougaard Nielsen, Ellen Løkkegaard, Anne Helms Andreasen, Yrsa Andersen Hundrup, and Niels Keiding. Estimating the effect of current, previous and never use of drugs in studies based on prescription registries. *Pharmacoepidemiology and drug safety*, 18(2):147–153, 2009.

[37] Lars Hougaard Nielsen, Ellen Løkkegaard, Anne Helms Andreasen, and Niels Keiding. Using prescription registries to define continuous drug use: how to fill gaps between prescriptions. *Pharmacoepidemiology and drug safety*, 17(4):384–388, 2008.

[38] American Association of Clinical Endocrinologist. *American Association of Clinical Endocrinologist Medical Guidelines for Clinical Practice for Developing a Diabetes Mellitus Comprehensive Care Plan*, volume 17. Endocrine Practice, 2011.

[39] PennState Eberly College of Science. Lesson18.3 kendall tau-b correlation coefficient.
https://onlinecourses.science.psu.edu/stat509/node/158.

[40] World Health Organization. The global burden of chronic.
http://www.who.int/nutrition/topics/2_background/en/.

[41] World Health Organization. Infographic: Ncd action plan.
http://www.who.int/nmh/publications/ncd-infographic-2014.pdf.

[42] Panagiotis Papapetrou, George Kollios, Stan Sclaroff, and Dimitrios Gunopulos. Discovering frequent arrangements of temporal intervals. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

[43] Manjiri Pawaskar, Machaon Bonafede, Barbara Johnson, Robert Fowler, Gregory Lenhart, and Byron Hoogwerf. Medication utilization patterns among type 2 diabetes patients initiating exenatide bid or insulin glargine: a retrospective database study. *BMC endocrine disorders*, 13(1):20, 2013.

[44] Jian Pei, Jiawei Han, and Wei Wang. Constraint-based sequential pattern mining: the pattern-growth methods. *Journal of Intelligent Information Systems*, 28(2):133–160, 2007.

[45] Anton Pottegård and Jesper Hallas. Assigning exposure duration to single prescriptions by use of the waiting time distribution. *Pharmacoepidemiology and drug safety*, 22(8):803–809, 2013.

[46] Lauren R Rodgers, Michael N Weedon, William E Henley, Andrew T Hattersley, and Beverley M Shields. Cohort profile for the mastermind study: using the clinical practice research datalink (cprd) to investigate stratification of response to treatment in patients with type 2 diabetes. *BMJ open*, 7(10):e017989, 2017.

[47] Allison B Rosen and David M Cutler. Challenges in building disease-based national health accounts. *Medical care*, 47(7 Suppl 1):S7, 2009.

[48] John Scott. *Social network analysis.* Sage, 2017.

[49] F Sjoqvist and Donald Birkett. Drug utilization. *Introduction to Drug Utilization Research.(WHO booklet) New York: WHO office of publications*, pages 76–84, 2003.

[50] American Diabetes Society. *Standards of Medical Care in Diabetes - 2015.* American Diabetes Society, 2012.

[51] Japan Diabetes Society. *Treatment Guide for Diabetes 2012-2013.* Japan Diabetes Society, 2012.

[52] W Stevenson. Kendall-tau, September 2012. http://statistical-research.com/wp-content/uploads/2012/09/kendall-tau1.pdf.

[53] Massoud Toussi, Jean-Baptiste Lamy, Philippe Le Toumelin, and Alain Venot. Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes. *BMC medical informatics and decision making*, 9(1):28, 2009.

[54] Yu Wang, Pengfei Li, Yu Tian, Jing-jing Ren, and Jing-song Li. A shared decision making system for diabetes medication choice utilizing electronic health record data. *IEEE Journal of Biomedical and Health Informatics*, 2016.

[55] Aileen P Wright, Adam T Wright, Allison B McCoy, and Dean F Sittig. The use of sequential pattern mining to predict next prescribed medications. *Journal of biomedical informatics*, 53:73–80, 2015.

[56] Mohammed J Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine learning*, 42(1):31–60, 2001.

Selected List of Publications

- **Journals**

[1] Khotimah, P.H. and Sugiyama, Y., Yoshikawa, M., Hamasaki, A., Sugiyama, O., Okamoto, K., Kuroda, T. Medication Episode Construction Framework for Retrospective Database Analyses of Patients with Chronic Diseases. *IEEE Journal of Biomedical and Health Informatics*, doi: 10.1109/JBHI.2017.2786741, 25 December 2017.

[2] Khotimah, P.H., Sugiyama, Y., Yoshikawa, M., Hamasaki, A., Sugiyama, O., Okamoto, K., Kuroda, T. Analyses of Physicians Medication Strategy for Type 2 Diabetes Treatments: Sequence Pattern Mining on Chronic Medication History. _*In Preparation_

- **International Conferences and Workshops**

[3] Khotimah, P.H. and Sugiyama, Y., Yoshikawa, M., Hamasaki, A., Okamoto, K., Kuroda, T. Revealing oral medication patterns from reconstructed long-term medication history of type 2 diabetes. In *In Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the (pp. 5599-5603). IEEE.*, doi: 10.1109/EMBC.2016.7591996.

[4] Khotimah, P.H. and Yoshikawa, M., Hamasaki, A., Sugiyama, O., Okamoto, K., Kuroda, T. Comparing frequent patterns: A study case of Apriori and singleton implementations in a diabetes type 2 data set. In *In Computer, Control, Informatics and its Applications (IC3INA), 2016 International Conference on (pp. 163-168). IEEE.*, doi: 10.1109/IC3INA.2016.7863043.