Medication Episode Construction Framework for Retrospective Database Analyses of Patients With Chronic Diseases

Purnomo Husnul Khotimah, Yuichi Sugiyama, Masatoshi Yoshikawa, Akihiro Hamasaki, Osamu Sugiyama, Kazuya Okamoto, and Tomohiro Kuroda, *Member, IEEE*

***Abstract*—*Objective*: For chronic diseases, medical history reconstruction is essential for retrospective database analyses. One important aspect is determining which prescriptions belong to the same episode. However, a standard framework for this task is still lacking, particularly for multitherapy datasets. This paper presents a medication episode construction framework for the medical history of patients with chronic diseases. *Methods*: Allen's relaxed temporal relations (i.e., temporal relations with time constraints relaxed by $\pm\epsilon$) is used to define the consecutive prescription relations considering the patients' behavior. For example, patients occasionally arrive earlier or later than their appointment. *Results*: ϵ influences the generation of stable periods (i.e., periods of time, at least three months, in which a medication is continuously taken by a patient). When using the lowest selected ϵ value (7 days), considerably fewer shorter stable periods (for durations less than 300 days) are produced and more longer stable periods are produced compared to cases without using ϵ. Furthermore, the results show that by using ϵ, regarding the number of events, where a stable period continues the previous stable period, decreases and the number of medication transition events available to be observed increases. *Conclusion*: Using ϵ in medication episode construction from multitherapy prescription datasets enables the longer expression of short-duration fragmented prescriptions and pruning repetitive prescriptions. *Significance*: Our proposed framework is designed for multitherapy datasets, which has not been addressed by previous studies. The concept of ϵ relaxes the prescription relation against noise caused by the patient behavior and consequently provides a compact, but informative search space for observing medication transition events in a longitudinal analysis.**

***Index Terms*—Medication episode construction, multitherapy dataset, chronic diseases.**

I. Introduction

PRESCRIPTION registries not only show a patient's medical history but can also be used as information sources for drug utilization and pharmacoepidemiology analyses [1]. Many of the studies that use prescription registries require the construction of treatment episodes [2]. One important aspect in treatment episode construction is determining which prescriptions belong to the same episode [3]. The process of reconstructing medical histories from prescriptions into other forms, such as treatment episodes, needs to be considered carefully because once the process is complete, the outcomes of subsequent activities will be based on the extracted data. However, previous studies on the use of prescription datasets only briefly discussed medical history reconstruction. Other studies discussed treatment episode construction with a focus on estimating drug exposure because the datasets do not include duration data. Hence, a standard framework for performing medical history reconstruction from prescription datasets is still lacking.

For chronic diseases, clinicians are often required to perform longitudinal analyses of medical histories over the years. For example, a common chronic disease is type 2 diabetes mellitus (T2DM). The treatment for such diseases is recommended to be patient centered, that is, respectful of and responsive to individual patient preferences, needs, and values [4]. In addition, the treatment spans for 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 [5]. Hence, for chronic diseases, assessing the physician's long-term strategy is necessary. Our research specifically focuses on observing medication transition events, which are when the medication treatment changes from one medication to the next. Medication transitions are important not only for revealing the physicians' actions toward the disease progress (changes in the patient's condition) but also for demonstrating the treatment development as new medicines or techniques are released [6].

A framework for medical history reconstruction is required for identifying of changes in medication based on hospital prescriptions. This is because there are several issues related to the

Manuscript received July 28, 2017; revised November 21, 2017; accepted December 15, 2017. Date of publication December 25, 2017; date of current version October 15, 2018. This work was supported by JSPS KAKENHI Grants 15H02705 and 15K00466. The work of P. H. Khotimah was supported by Indonesian Endowment Fund for Education. *(Corresponding author: Masatoshi Yoshikawa.)*

P. H. Khotimah, Y. Sugiyama, and M. Yoshikawa are with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: hkhotimah@db.soc.i.kyoto-u.ac.jp; oogfranz@gmail.com; yoshikawa@i.kyoto-u.ac.jp).

A. Hamasaki is with the Center for Diabetes & Endocrinology, Taduke Kofukai Medical Research Institute, Kitano Hospital, Osaka 530-8480, Japan (e-mail: hamasaki@kuhp.kyoto-u.ac.jp).

O. Sugiyama is with the Preemptive Medicine & Lifestyle-Related Disease Research Center, Kyoto University Hospital, Kyoto 606-8507, Japan (e-mail: sugiyama@kuhp.kyoto-u.ac.jp).

K. Okamoto and T. Kuroda are with the Division of Medical Information Technology and Administration Planning, Kyoto University Hospital, Kyoto 606-8507, Japan (e-mail: kazuya@kuhp.kyoto-u.ac.jp).

Digital Object Identifier 10.1109/JBHI.2017.2786741

nature of prescription datasets. First, the periods of two consecutive prescriptions are occasionally unconnected and overlap with each other because a patient may arrive earlier or later than their appointment. Second, prescriptions generally have short durations as a result of 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 must be able to express fragmented prescriptions in aligned medication episodes (i.e., a period of time when a doctor prescribes the same medication continuously) such that we can observe the medication transitions precisely.

Our next concern in prescription reconstruction is 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 because patients that have more than one medication at a time are excluded in monotherapy datasets. However, in multitherapy datasets, there can be more than one prescription at one time, and there can be different medications than those currently taken by the patient. Consequently, the prescriptions can have more possible temporal relations. For example, when a prescription has not been finished, a patient may visit a physician and receive a new short duration of the prescription that does not last as long as the previous on-going prescription. This event will lead to the 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 [2] attempted to construct the treatment episode of an antidepressant treatment. The study compared two treatment episode construction methods using a prescription time gap parameter. The first method does not add the 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 (ATC) classification system. A treatment episode constructed by both methods in [2] is a period of time that 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 having the same medication) or there can be changes in the medications (i.e., one treatment episode having more than one medication). This situation is different from our goal of reconstructing medical histories through constructing medication episodes from a prescription dataset. In addition, [2] only considered monotherapy by excluding patients who are prescribed more than one medication at one time. This situation is in contrast to the nature of medication for chronic diseases, which includes multitherapy prescriptions. In our study, the proposed method takes a multitherapy dataset into account by using possible temporal relations between consecutive events that had been defined by Allen in [7] to address more possible prescription relations available in multitherapy datasets. In addition, to address situations where patients may visit the hospital earlier or later than their appointment, we have employed the concept of time margin (ϵ) to provide flexibility in assigning the temporal relation between consecutive prescriptions. As we will show in a later section, this variable is important for the medication episode construction in chronic disease analyses.

Furthermore, for diabetes treatment, the physician often needs to wait three months to evaluate the effectiveness of the medication. The recommended waiting time is three months which we refer to as the *3 month rule* [5]. Based on the *3 month rule*, we hypothesize that medication episodes that have a duration of at least three months and those with longer durations have more essential meaning for longitudinal analyses. This hypothesis is because a longer duration of medication indicates that the patient's condition is reaching the control target assigned by the physician. Thus, we need to be able to observe the medication transition events between these types of medication episodes, and we later refer to this medication episode as a stable period. Hence,our proposed medication episode reconstruction is also for enabling the identification of such stable periods, which has not been discussed in previous related studies.

This paper presents a medication episode construction framework for chronic diseases and it is applied to a multitherapy hospital prescription dataset of T2DM patients. Specifically, we emphasize the use of Allen's temporal relations and the time error margin in the framework, as well as the identification of stable periods. Our main contributions are as follows:

- In our method, we use the temporal relation defined by Allen [7] to accommodate the possible prescription relations in a multitherapy dataset.
- We use the concept of epsilon (time error margin) to relax the prescription temporal relation considering the noise caused by the patient's behavior or the physician's behavior, which is essential for expressing the fragmented short-duration prescription in a medication episode with a longer duration.
- We use the time period that is recommended for assessing the effectiveness of the medications to identify the stable periods.

The preliminary results of the study of T2DM patients' long-term medical history when treated with oral medication have been reported in [8]. [8] explained the general idea of the method to search for frequent patterns from a longitudinal prescription registry dataset and other related clinical data. Medication episode construction was introduced as one of the steps in the method. Moreover, the results reported in [8] focused mainly on the unfamiliar patterns found by the methodology. Conversely, this paper focuses on the medication episode construction framework and analyzes the data behavior according to the selection of the ϵ value.

The remainder of this paper is organized as follows. In Section II, we introduce the theoretical background used as the foundation for our study. In Section III, we present the proposed framework for constructing medication episodes from multitherapy prescription datasets. In Section IV, we describe our experimental results. Section V discusses the significance of ϵ in the framework. Finally, we conclude the paper in Section VI.

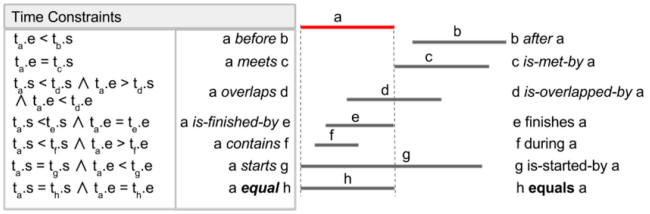

Fig. 1. Allen's temporal relations with time constraints.

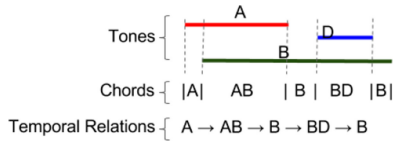

Fig. 2. Morchen's chords.

II. Background

A. Temporal Relations of Interval-Based Data

Our proposed framework for constructing medication episode is closely related to temporal relations between interval-based events. The first to address this topic was Allen in [7]. [7] introduced seven temporal relations ($before$, $meets$, $overlaps$, $is-finished-by$, $contains$, $starts$, and $equal$) with their inverses (13 relations in total) as shown in Fig. 1. Temporal relations between interval-based events enable capturing or extracting temporal knowledge from natural language or relative information, for example, increasing medication dosage $after$ the A1c value increases.

Allen's temporal relations have been widely used and developed. [9] added end time points (start and end times) of the interval to define the interval constraint relations. Fig. 1 presents the time constraints for each temporal relation. In [9], only seven temporal relations (the left part of Fig. 1) are used such that one event is represented by only one temporal relation to avoid confusion. [10] used a temporal relation matrix to capture all possible relations in multivariate series datasets. The size of the matrix increased with the number of observed intervals. [11] introduced the concept of time error margin (ϵ) to define more flexible matches between two interval-based events. Hence, the temporal relations, which were previously strictly constrained by the 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 incidents: event A has no gap from event B and event C has a 1 day gap from event D. Hence, by using an 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, [12] used the Karma Lego algorithm to find temporal interval relation patterns between the A1C level and the defined daily dosage (DDD) of diabetes medication. In our case, we use the temporal relations as a tool in the medication episode construction to determine which rule should be applied for different prescription relations.

Furthermore, our framework is related to Morchen's time series knowledge representation (TSKR) model, which was proposed in [13]. Their study showed that Allen's relations are not robust for noisy time series, ambiguous because one relationship may represent different conditions, and not easily comprehensible because one condition may be represented by different relationships [13]. TSKR was proposed to mine multivariate time series data by transforming the time series into interval symbolic series and finding the coinciding intervals. Morchen proposed the interval series as $tones$ (i.e., observed parameters) and the interval coincide series as $chords$. In Fig. 2, we are able to observe three tones (A, B, and D). These three tones result in four chords (A, AB, B, BD and B). The temporal relations of these chords will be represented as A → AB → B → BD → B. Such a representation is suitable, particularly for studies that observe more than two parameters because the number of all possible Allen's temporal relations will highly increase as the number of observed parameters increases.

Hence, to observe medication transition events, the $chords$ model can be used to represent the overlapping prescriptions prescribed by the physicians as the medication episode.

B. Medical Episode Construction

In pharmacoepidemiology and drug analysis studies, patients' drug episodes are often assessed. The prescription registry as a data source often omits the duration. The available information generally includes the date of redeeming the prescription and the amount of dispensed medication. Hence, approximating patients' actual drug use is needed. Several methods to estimate the duration of each prescription have been discussed in [2], [3], [14]–[16].

The first method consists of using the ratio of prevalence and incidence rate and using this ratio as a constant period for each prescription. This is performed by assuming a constant use of dose (e.g., DDD) or assuming other fixed amounts [14]. However, this method has been reported to have several caveats [14]. The second method is by using the waiting time distribution (WTD) [3], [14]. The WTD is a frequency distribution of the 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 [2], [15], [16]. By using the redemption date and the dispensed amount, it is possible to define drug use episodes. However, determining whether and when the dispensed drugs are used is difficult.

Thus far, the aforementioned methods focused on estimating the duration rather than reconstructing the prescription. However, as mentioned in the previous section, our main idea of connecting prescriptions is the most similar to [2], where prescriptions were concatenated 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 for treatment episode construction based on the maximal gap were introduced, as shown in Fig. 3. The first

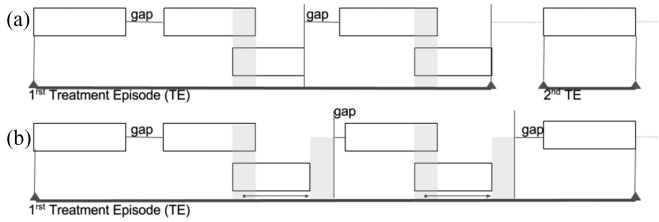

Fig. 3. Two methods introduced in [2].

method shown in Fig. 3(a) does not add the overlap duration at the end of the expected end time of the treatment episode, whereas the second method adds the overlap duration when the medication belongs to the same ATC group because the patient may arrive earlier as in Fig. 3(b). Hence, the second method causes the original gap to become shorter. Both of the methods introduced in [2] were applied for a monotherapy dataset; consequently the considered prescriptions were prescriptions that next to each other either have a gap or not, and overlap with each other. In addition, the study in [2] focused on comparing the effect of maximal gap variation on both methods.

The situation is different for our study case, in which we use a multitherapy dataset. Hence, there are more possible temporal relations between consecutive prescriptions compared to a monotherapy dataset. Moreover, we consider the fact that a patient may arrive earlier or later than their scheduled appointment, which means that successive prescriptions with a 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 identifying short gaps from longer gaps but also identifying short overlaps from longer overlaps. Compared to the two methods introduced by [2], our method appears to be similar to the second method [2] (the one that adds the overlap duration). In the second method [2], the successive prescriptions are considered to be overlapping irrespective of the overlap interval (whether it is a short or long overlap). However, in our method, the prescription relations between two successive prescriptions are relaxed by $\pm\epsilon$. Hence, the successive prescriptions are considered to be overlapping when the overlap duration is more than ϵ. Furthermore, the focus of our study is different from that of [2], where the emphasis was comparing two construction methods (not adding overlap and adding overlap). Our study focuses on the ϵ variation in the generation of stable periods to enable observing the medication transition events. To summarize, Table I compares the properties between [2] and our study.

TABLE I
COMPARISON OF THE PROPERTIES IN OUR STUDY WITH [2]

Property	Related previous study [2]	Our Study
Dataset	Monotherapy	Multitherapy
Parameter used	Maximum 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 transformed into a $meets$ relation	Prescriptions with a gap that is not more than a predefined parameter value	Prescriptions with a gap or overlap that is not more than a 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

C. Type 2 Diabetes Mellitus (T2DM)

To provide an overview of the role of retrospective-database analyses for chronic diseases, we use the case of diabetes because it is a common chronic disease. Diabetes is a progressive chronic disease, and it may lead to increases in risk factors for other conditions, such as heart disease, amputation, and kidney failure. There two types of diabetes: type 1 and type 2 diabetes. Type 2 diabetes mellitus (T2DM) receives considerable attention due to its high prevalence at the global scale. T2DM is caused by the patient's body being unable to properly use insulin, which is called insulin resistance. Initially, the pancreas produces extra insulin to compensate for insulin resistance. Over time, however, the pancreas is not able to keep up and cannot secrete sufficient insulin to maintain normal blood glucose levels.

As a chronic and progressive disease, T2DM needs to be managed in a comprehensive and longitudinal manner. Medical societies currently publish medical guidelines for providing T2DM stakeholders with standard and evidence-based recommendations. To develop such medical guidelines, clinicians are required to perform retrospective studies. Medical history recorded as a prescription dataset is a good resource [17]. However, the current method for finding temporal patterns from the raw dataset (prescriptions) is performed without first constructing the medication episode, identifying the stable periods, or even using the common method as in [12], [17]. As we previously mentioned, T2DM treatment continues for years, and the effects of the medications need to be assessed over a long period of time. However, prescriptions are fragmented, repetitive and progressive over the years. Hence, medication episode construction is necessary prior to conducting a longitudinal analysis. In addition, the related studies regarding medication episode construction are focused on monotherapy. This is in contrast to the medical guidelines, which recommend that T2DM pharmacotherapy begins with diet therapy and monotherapy and then proceeds to dual therapy, triple therapy, combination therapy with injection drugs, or switching to insulin therapy [5]. Therefore, a longitudinal and multitherapy analysis could provide an exhaustive analysis tool for clinicians to conduct retrospective studies.

III. METHODOLOGY

In this section, we describe the proposed medication episode construction framework for enabling medication transition

TABLE II
MEDICATION TYPES

No	Medication Type	Medication Names
1	Sulfonylurea (SU)	Rastinon, Euglucon, 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

events between stable periods. The description includes the input (the prescriptions), the method (medication episode construction), and the output (the stable period identification).

A. Prescription

Diabetes medications are classified into several types depending on their mechanism of action. Table II shows medication types with their names.

Definition 1: Medication name **medName** is the proprietary name of the medication. Each medication name belongs to a single medication type **medType** (i.e., medication classification based on the mechanism of action).

Definition 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[1], $m[]$ array of medication label, and $d[]$ array of medication dosages w.r.t. the medication label. Medication label can be a medName or medType. The dataset provides the start time and duration dur, and e is obtained from $s + dur$. A prescription dataset is a sequence of prescriptions $[P_1, P_2, ..., P_n]$, where prescriptions are ordered by the start 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 III 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.

B. Prescription Relation

The prescription relation represents the temporal relation between prescriptions in a time line. Fig. 4 shows the possible temporal relations between consecutive prescriptions, for example, P_1 $before$ P_2, P_8 $starts$ P_9, and P_9 $overlaps$ P_{10}.

[1]In Japan, it is mandatory, by regulation, to provide duration information for the prescriptions.

TABLE III
FULL PRESCRIPTION DATASET

P_n	pid	did	s	e	m[]	d[]
P_1	7	1	1	56	A	d_a
P_2	7	1	59	125	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	340	375	A	d_a
P_6	7	1	380	390	B	d_b
P_7	7	1	397	407	A	d_b
P_8	7	1	407	420	A	d_a
P_9	7	1	407	420	B	d_b
P_{10}	7	1	421	443	A	d_a
P_{11}	7	1	426	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	755	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

P_n : Prescription id; pid: patient id; did: doctor id; s: start time; e: end time; m[]: array of medication label; d[]: array of dosage w.r.t. the medication label.

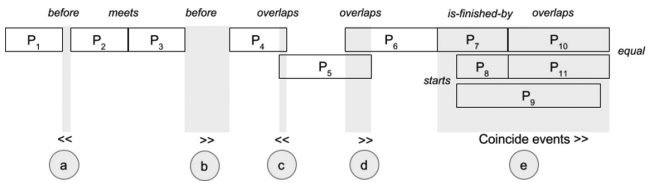

Fig. 4. Allens temporal relations in aligned prescription sequence.

As shown in Fig. 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**. This case is similar with event **c** and event **d** (i.e., the overlap duration on event **c** is considerably smaller than that on event **d**). In medication episode construction, such conditions may require different treatments. For example, the fact that a patient may arrive earlier or later than their appointment may cause short gaps and overlaps and should be treated as a $meets$ prescription. Meanwhile, a longer gap and overlap should be treated without modification. However, the maximum gap in medication episode construction will only influence prescriptions with a gap ($before$ relation). For overlapping prescriptions, irrespective of how small the overlaps are, the duration will be treated as overlaps. In this situation, the time error margin (ϵ) is suitable for assigning the prescription relation.

Definition 3: Epsilon, ϵ, is a 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:

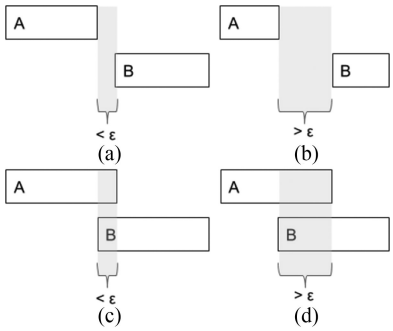

Fig. 5. Time error margin (ϵ).

$$t_1 =_\epsilon t_2 \leftrightarrow |\, t_1 - t_2 \,| \leq \epsilon$$
$$t_1 <_\epsilon t_2 \leftrightarrow 0 < t_2 - t_1 > \epsilon$$

Example 1: Based on Fig. 5, if we use the concept of ϵ, then the prescription relations in Fig. 5(a) and (c), which were previously A $before$ B and A $overlaps$ B, will be A $meets$ B. For Fig. 5(b) and (d), the prescription relation will remain the same.

To demonstrate the prescription relations from the raw data, Fig. 6 shows the prescription diagram of Table III. The x-axis shows the time per 30 days. Fig. 6(a) shows the aligned prescriptions' durations based on their start time, end time and medication label. As shown in Fig. 6(a), a prescription can have a short duration, and it occasionally overlaps with other prescriptions or has a gap. As previously mentioned, this situation may occur because a patient may arrive earlier or later than their appointment with the physician. Another event that we observe is that many of the prescriptions continue the previous medication. Furthermore, the prescriptions begin overlapping with each other when the prescription is modified by the physician. For example, with the transition from medication A to dual-therapy AB at approximately $time = 13$, we have an overlapping prescription between medications A and B. Another example is shown when the medication is changed from dual-therapy AB to ABC at approximately $time = 24$ and when the medication is switched from ABC to ABD at approximately $time = 32$.

Example 2: Using P_1, P_2, and P_3 from Table I and $\epsilon = 14$ days, we have the following two prescription relations:

$$|P_2.s - P_1.e| < \epsilon \Rightarrow P_1\ meets\ P_2$$
$$|P_3.s - P_2.e| > \epsilon \Rightarrow P_3\ before\ P_2$$

C. Medication Episode

To reconstruct a continuous medication episode, as in Fig. 6(b), we use Allen's temporal relations relaxed by the concept of ϵ, as shown in Fig. 7. Our main 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 medication from the hospital, and in such cases, the physician adjusts the medications based on the patient's condition. For equal prescriptions, that is, prescriptions with the same start time and end time, we merge the prescriptions. Further details on the rules of medication episode construction are explained in the previous publication [8].

Definition 4: A medication episode ME is a concatenation of $meets$ prescriptions that have the same medication label and dosage. ME shows the period of time when the physician does not change the prescription. The ME dataset is $\mathbf{ME} = \{\, \mathrm{ME}_1, \mathrm{ME}_2, .., \mathrm{ME}_n \,\}$, where n is the total number of MEs in the patient's medical history and ME_n is ordered based on the start time and end time.

Example 3: From Table III, $P_1\ meets\ P_2$, $P_1.m[] = P_2.m[]$, and $P_1.d[] = P_2.d[]$. Hence, P_1 and P_2 are concatenated into a single ME.

Recalling Fig. 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 Fig. 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 5: Threshold δ is the minimum period, in days, for the physician to observe the medications effectiveness.

Definition 6: A stable period SP is a medication episode in which the duration is at least equal to δ days. It is defined as $SP = \{SP \in \mathbf{ME} | SP.e - SP.s \geq \delta\}$.

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

Definition 7: A trial/short period TP is a medication episode where the duration is less than the threshold δ days. It is defined as $TP = \{TP \in \mathbf{ME} | TP.e - TP.s < \delta\}$.

Definition 8: An unstable period UP, is a single TP or an aggregation of consecutive TPs.

Definition 9: A blank period BP, is a period of time in which there was no medication recorded in the medical history after ϵ days.

D. Medication Transition Events

After identifying the SP, we obtain the SP sequence. From Fig. 6(c), we have an SP sequence of $A \rightarrow A \rightarrow AB \rightarrow AB \rightarrow ABC \rightarrow ABD$. We define a 1-consecutive sequence as a single sequence between two consecutive SPs, for example, A $\rightarrow$ A, A $\rightarrow$ AB, AB $\rightarrow$ AB, AB $\rightarrow$ ABC, and ABC $\rightarrow$ ABD. Medication transition events occurred on transition points (i.e., the point between consecutive SPs where medication transition event(s) occur). Based on the SP sequence from Fig. 6(c), transition points are the points between $A \rightarrow AB$, $AB \rightarrow ABC$, and $ABC \rightarrow ABD$. For medication transition events, we list the following five:

- Add is when new medication(s) are added to the previous medication.
- $Stop$ is when previous medication(s) are stopped from the previous medication.
- $Switch$ is when new medication(s) are added and previous medication(s) are stopped.

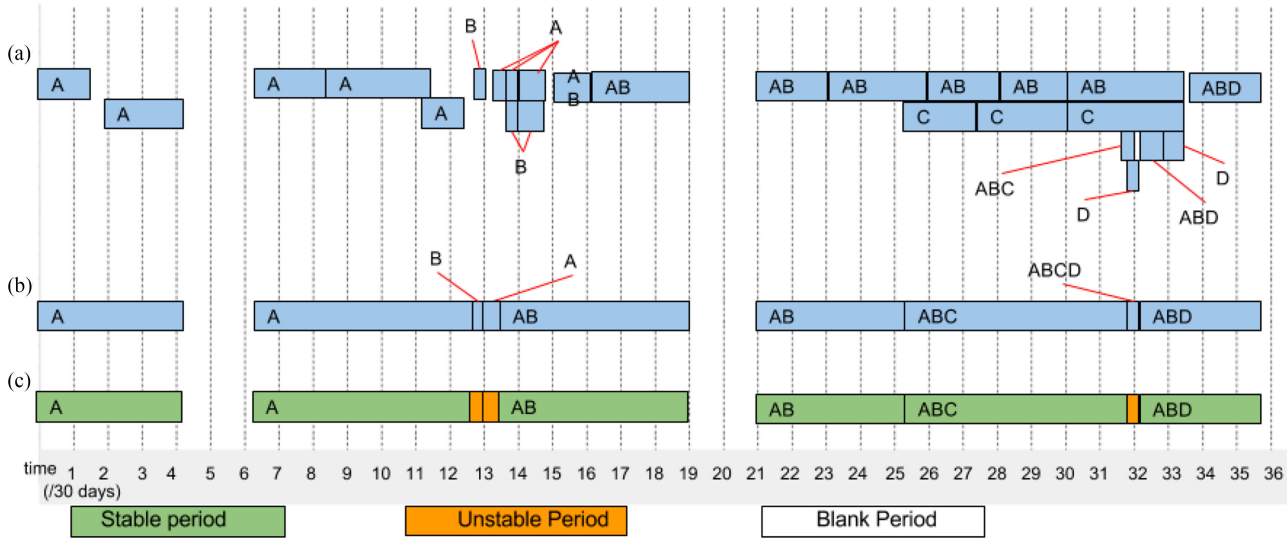

Fig. 6. Physician's prescription diagram as listed in Table III.

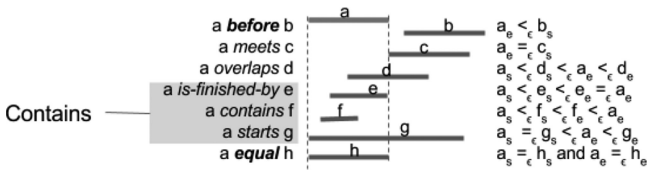

Fig. 7. Allen's left-side temporal relation using epsilon.

- *Increasing* is when the dosage of medication(s) is increased.
- *Decreasing* is when the dosage of medication(s) is decreased.

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

Example 4: Based on Fig. 6(c), we consider $A \rightarrow AB$ and $AB \rightarrow ABC$ as *add* events and $ABC \rightarrow ABD$ as a *switch* event.

IV. Experimental

The investigation in this paper focuses on three aspects that support our main contributions, as follows: the nature of multitherapy datasets, the significance of ϵ in connecting successive prescriptions, and the generation of SPs affected by variations in the ϵ value.

A. 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 prescriptions. The prescriptions are extracted between September 2000 and August 2015 for the medications listed in Table II.

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

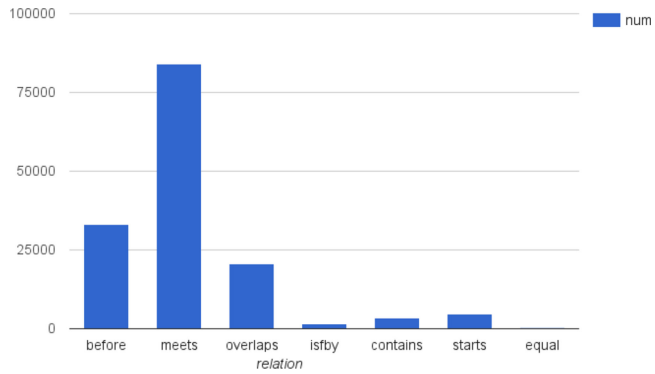

Fig. 8. Number of each prescription relation extracted based on the fixed Allen's relations.

B. Results

First, to show the nature of the multitherapy dataset, we extracted the numbers of prescription relations based on Allen's relations. Fig. 8 shows the number of each prescription relation: $before$, $meets$, $overlaps$, $is-finished-by$ (isfby), $contains$, $starts$, and $equal$. This figure 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 that have the same time range given by a different physician as defined in the full prescription.

Second, to be able to observe how much of the $before$ and $overlaps$ portions will be affected by the use of ϵ in the construction, we investigated the numbers of prescriptions with $before$ and $overlaps$ relations that have a gap and overlap $\leq$ than the ϵ value variation. In this investigation, we varied the ϵ value based on our initial assumptions that a patient may arrive earlier or later than their appointment. The values that represent our assumptions are approximately one week (7 days), two weeks (14 days), and three weeks (21 days). The results are presented in Figs. 9 and 10. As shown in Fig. 9, more than 30% of the

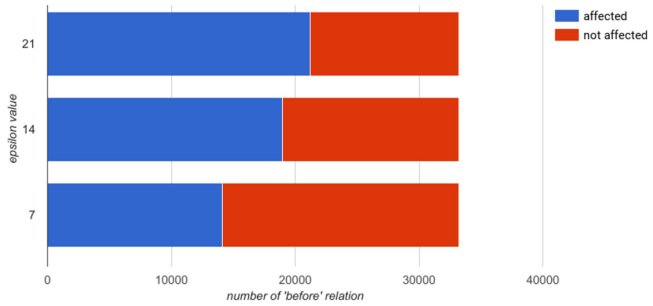

Fig. 9. Prescription number with $before$ relations.

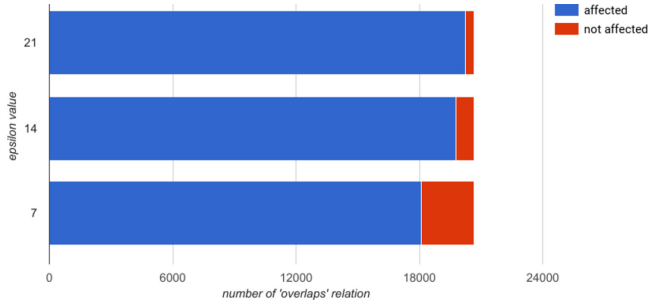

Fig. 10. Prescription number with $overlaps$ relations.

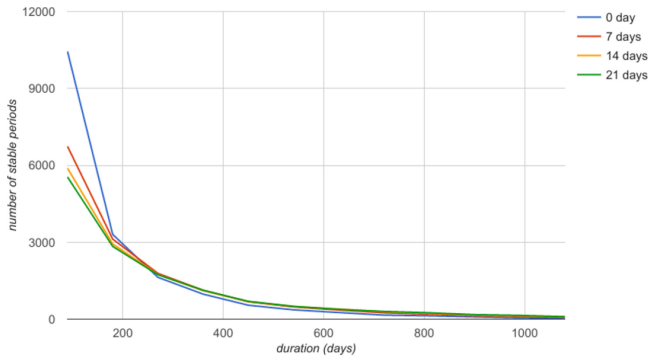

Fig. 11. Number of stable periods with length duration.

$before$ prescriptions were affected using the smallest value of ϵ (7 days). Additionally, the number of affected prescriptions increases with the increasing ϵ value. As shown in Fig. 10, more than 80% of the $overlaps$ prescriptions are affected by the ϵ value.

We also investigated the number of generated SPs that have a certain duration upon a variation in the value of ϵ, as presented in Fig. 11. In this experiment, we use 90 days as the value of δ based on the 3 $month\ rule$ and we construct the medication episode by using the $medName$ as the medication label. The lines display the accumulated number of SPs. The x-axis represents the duration per 90 days. The blue line represents the number of SPs generated by the construction without ϵ ($\epsilon = 0$). The red, orange, and green lines are generated by the construction with ϵ values of 7 days, 14 days, and 21 days, respectively. As shown in this figure, the number of SPs with a duration of less than 200 days sharply decreases when we use the concept of ϵ (red, yellow, and green lines). The blue line (without ϵ) has a higher number of short SPs compared to the other lines (under

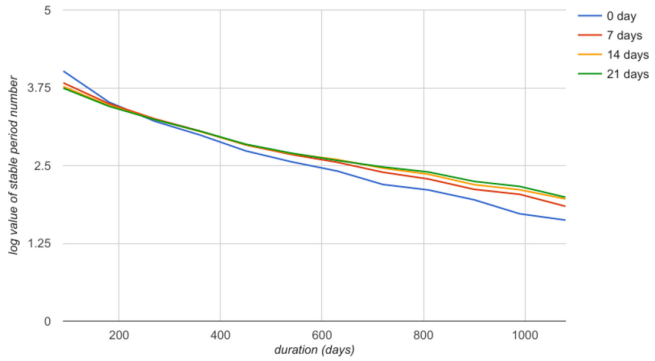

Fig. 12. Log scale from Fig. 11.

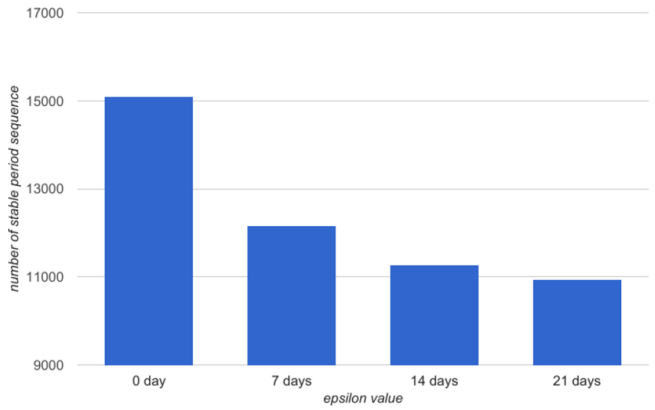

Fig. 13. Number of 1-consecutive sequences based on the ϵ value.

200 days). Furthermore, the blue line produces less SPs with duration longer than approximately 300 days.

To observe the discrepancy from each line for durations of more than 300 days, we also generated the log scale of Fig. 11, which is displayed in Fig. 12. Fig. 12 shows that the blue line falls (construction without ϵ) under the other lines for SPs with a duration of more than 300 days. Another observation from both figures is that the SPs generated using ϵ have only slight differences.

A deeper observation on the number of 1-step-sequence out of SPsequences from the constructed medication episode is presented in Fig. 13. In Fig. 13, the disparity between the ϵ value selection is clearly observable. The number of SP sequences is decreasing with higher ϵ values. A further observation between the number of continue patterns and transition events in the stable period sequence is possible from Fig. 14. Fig. 14 shows that the number of continue events significantly decreases with increasing ϵ value. In contrast, the number of transition events available to be observed increases with increasing ϵ value. Note that the total number between continue events and transition events is not the same as the number of 1-consecutive sequences. This difference is because there can be more than one transition event in one transition point.

V. Discussion

In this section, we discuss the results presented in the previous section. As shown in Fig. 8, the prescription relations in a

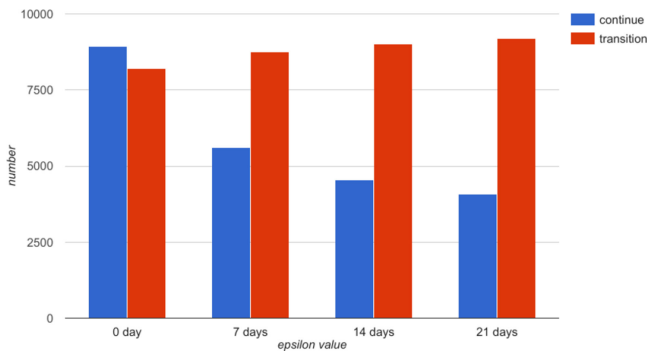

Fig. 14. Number of continue events and transition events based on the ϵ value.

multitherapy dataset cover all Allen's relations. The high number of the $meets$ relation indicates that it is common for chronic patients' prescriptions to be connected to each other, which is caused by either the patient's behavior (to arrive on time) or the physician's need to separate continuous medication due to regulations, such as medical doctors cannot prescribe medications longer than a certain period of time. When the $meets$ prescriptions have the same medication, they will be concatenated and form longer medication episode. Hence, the medication episode construction is still able to generate longer medication episodes even without using ϵ. However, the numbers of $before$ and $overlaps$ relations are also significant in a multitherapy dataset. Hence, incidents in which patients arrive earlier or later than their scheduled appointment is also common. This patient behavior of arriving earlier than their scheduled appointment may occur when a patient is unable to arrive on time because of other reasons and then decided to arrive earlier to renew the prescription. In this situation, the overlap duration is considered to be short. However, in patients with chronic diseases, it is possible for patients to have changes in conditions that cause them to arrive earlier than their scheduled appointment. This situation may have a longer overlap duration. The patient behavior of arriving later than their scheduled appointment may occur when a patient decided to come later because they were unable to arrive on schedule. In this situation, the gap duration is generally a short duration because for patients with chronic diseases, it is important for them to take the medication. Another situation may cause the gaps, that is, because the patient visits a private clinic or other health care provider. This situation generally causes a longer gap. Hence, we need to be able to identify which of these $before$ and $overlaps$ prescriptions have a short gap or overlap duration. Thus, we need to consider such incidents in the construction and develop the rules for the medication episode.

As shown by the results present in Figs. 9 and 10, the proportions of prescriptions with $before$ and $overlaps$ relations that were affected by ϵ are significant. Using the concept of ϵ, the construction of the medication episode was 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 the previous method in [2] that used the maximum allowed gap, the method only affected prescriptions with a $before$ relation while the prescriptions with an $overlaps$ relation continued to have the same relation despite their short overlap durations. As shown in Fig. 10, there are prescriptions with short and longer overlap durations. If the method in [2] is used, the size variation of $overlaps$ will not be addressed because a fixed temporal relation is used. In this case, a short overlap duration will not be identified, and then it will be treated as an overlap and merged in the case where the medications from both successive prescriptions are different, which will later produce more short unstable periods.

In addition, there are considerably low numbers of $is-finished-by$, $contains$, $starts$ and $equal$ relations from Fig. 8. Although the numbers are low, these prescription relations have significant meaning because they represent incidents of patients with chronic diseases that occasionally occur when there are temporarily abrupt changes in the 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 the previous example (Fig. 6), which is when the patient's condition changes, medication transition events occur with prescriptions with a $contains$ ($is-finished-by$, $starts$, and $contains$) relation. Therefore, to retain the medication transition information, we need to address these temporal relations.

Regarding other technical aspects of the medication episode construction, [2] 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 in the value of the maximum allowed gap. As an addition to the discussion of our study, note that using the notion of ϵ ($\epsilon > 0$) also has an effect on the number of SPs as shown in Fig. 11. Without ϵ, the medication episode construction still generated SPs. The SPs generated without ϵ were produced in a very significant number for durations of less than 200 days. This result is in contrast to the SPs generated using ϵ, which is shown with the large gap between the blue line and the other lines in shorter durations (less than 200 days). This data behavior may be caused because many SPs generated without ϵ are merely a continuation of the previous SPs 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 arriving slightly later than the scheduled appointment. Therefore, the use of ϵ is significant in avoiding such situations. Moreover, based on the generated SPs as shown in Fig. 12, the selection of a higher ϵ value will have greater performance in producing SPs with longer durations (more than 300 days). This result occurs because the construction using ϵ will be able to connect prescriptions separated by short gaps, which will produce SPs with longer durations. Hence, ϵ is essential for producing longer expressions from prescriptions. Furthermore, this result shows that the number of SP transitions available for further analysis (search space) is affected by the selection of the ϵ value, and the search space size influences the cost of data-driven analyses [18].

Moreover, from Fig. 13, we are able to observe that the selection of the ϵ value also influences the number of 1-consecutive sequences out of the SPsequence in each patient. Further

observation of Fig. 14 shows that the number of continue events in the SPsequence is sharply decreased. This result confirms the previous statement that many of SPs are merely continuations from the previous one, which are then concatenated by ϵ. In contrast, the number of transition events is increasing, indicating that there are unstable periods connected with ϵ, which then construct an SP and more medication transition events that can be observed. Regarding this result, a study by [12] introduced a horizontal support value, which is the number of instances of the pattern found in an entity (e.g., a single patient's 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.

Based on the results in Figs. 9–14, additional insight can be obtained. The insight is that the results show saturation after ϵ equal to 14 days, which means that the results do not show a significant difference between ϵ values of 14 days and 21 days. A medical doctor from the Diabetes department of Kyoto University Hospital confirmed that this observation is consistent with the practice of the Diabetes department of Kyoto University Hospital, where the patients, in many cases, visit the hospital up to 14 days earlier or later than their appointment if they do not visit on the appointment day.

From a medical application perspective, clinicians currently assess the chronic disease medication from the medication history in the form of prescription datasets, which are difficult to use particularly for longitudinal analyses because of their characteristics (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 findings. As in diabetes, medication effect is commonly assessed in longer-duration observation windows. By observing the medication transition from one SP to the next, we are able to reveal an unfamiliar pattern that was possibly driven by a newly released medication (DPP4i) [8]. Other possible relevant applications that may benefit from the proposed framework are drug utilization or pharmacoepidemiology studies, for example, in studies that take duration as an essential factor to investigate: drug exposure and drug survival analyses (i.e., studies that assume that a drug that "survives" longer in treatment will be one that is safer and/or more effective).

Finally, the dataset 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 medication dosage or even if there was hospitalization because in the case of hospitalization, the patient is taking medication provided by the hospital only. The current assumption used in our rules is that coinciding medication will be merged (i.e., different medications with same medication type will be pruned as defined in the rule of $merging$). However, such conditions generally occur in a short time (less than 3 months). Hence, for longitudinal analyses, we are concerned with the medication transition events between SPs rather than the unstable period. In addition, this method is applicable for other groups of diseases, including HIV and periodontitis, which can be performed by customizing the epsilon value. For example, we need to set a new epsilon value for HIV. For periodontitis, the epsilon value should be shorter because it is an infection. However, our current study does not include datasets of patients with two interrelated chronic diseases, such as diabetes and periodontitis [19] or HIV and diabetes [20].

VI. Conclusion

This paper investigates the preparation of data for retrospective database analyses for observing medication transition events. To the best of our knowledge, there are no previous medication episode construction frameworks that incorporated all possible Allens temporal relations for multitherapy datasets. By accommodating Allen's relations in the ruled-based construction, we are able to preserve prescription information in a multitherapy dataset which would otherwise be missing. Furthermore, the use of ϵ in expressing Allen's relations is significant in reducing repetitive medication episodes, constructing higher numbers of longer medication episodes and enabling more medication transition events to be observed. This is important for longitudinal analyses of chronic diseases, particularly for observing the strategic actions by the physician to achieve an ideal condition for the patients. In a previous study, [2] 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, where the selection of the epsilon value influences the generation of SPs 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 value significantly affects the measurement of further analysis results.

References

[1] F. Sjoqvist and D. Birkett, "Drug utilization," *Introduction to Drug Utilization Research.* New York, NY, USA: WHO Office Publ., 2003, pp. 76–84.
[2] H. Gardarsdottir, P. C. Souverein, T. C. Egberts, and E. R. Heerdink, "Construction of drug treatment episodes from drug-dispensing histories is influenced by the gap length," *J. Clin. Epidemiol.*, vol. 63, no. 4, pp. 422–427, 2010.
[3] A. Pottegård and J. Hallas, "Assigning exposure duration to single prescriptions by use of the waiting time distribution," *Pharmacoepidemiol. Drug Safety*, vol. 22, no. 8, pp. 803–809, 2013.
[4] Y. Wang, P. Li, Y. Tian, J.-j. Ren, and J.-s. Li, "A shared decision making system for diabetes medication choice utilizing electronic health record data," *IEEE J. Biomedical Health Inf.*, vol. 21, no. 5, pp. 1280–1287, Sep. 2016.
[5] Japan Diabetes Society, "Treatment Guide for Diabetes 2012–2013," Jpn. Diabetes Soc., Tokyo, Japan, 2012.
[6] M. Pawaskar, M. Bonafede, B. Johnson, R. Fowler, G. Lenhart, and B. Hoogwerf, "Medication utilization patterns among type 2 diabetes patients initiating exenatide bid or insulin glargine: A retrospective database study," *BMC Endocrine Disorders*, vol. 13, no. 1, pp 13–20, 2013.
[7] J. F. Allen, "Maintaining knowledge about temporal intervals," *Commun. ACM*, vol. 26, no. 11, pp. 832–843, 1983.
[8] P. H. Khotimah, Y. Sugiyama, M. Yoshikawa, A. Hamasaki, K. Okamoto, and T. Kuroda, "Revealing oral medication patterns from reconstructed long-term medication history of type 2 diabetes," in *Proc. 2016 IEEE 38th Annu. Int. Conf., Eng. Med. Biology Soc.*, 2016, pp. 5599–5603.

[9] P.-s. Kam and A. W.-C. Fu, "Discovering temporal patterns for interval-based events," in *Proc. Int. Conf. Data Warehousing Knowl. Discovery*, 2000, pp. 317–326.
[10] F. Höppner, "Learning temporal rules from state sequences," in *Proc. IJCAI Workshop Learn. Temporal Spatial Data*, 2001, vol. 25, pp. 25–31.
[11] P. Papapetrou, G. Kollios, S. Sclaroff, and D. Gunopulos, "Discovering frequent arrangements of temporal intervals," in *Proc. IEEE 5th Int. Conf. Data Mining.*, 2005, pp. 354–361.
[12] R. Moskovitch and Y. Shahar, "Medical temporal-knowledge discovery via temporal abstraction," in *Proc. Amer. Med. Informat. Assoc.*, 2009, pp. 452–456.
[13] F. Mörchen, "A better tool than Allens relations for expressing temporal knowledge in interval data," in *Proc. Workshop Temporal Data Mining 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 25–34.
[14] J. Hallas and H. Støvring, "Templates for analysis of individual-level prescription data," *Basic Clin. Pharmacol. Toxicol.*, vol. 98, no. 3, pp. 260–265, 2006.
[15] L. H. Nielsen, E. Løkkegaard, A. H. Andreasen, and N. Keiding, "Using prescription registries to define continuous drug use: How to fill gaps between prescriptions," *Pharmacoepidemiol. Drug Safety*, vol. 17, no. 4, pp. 384–388, 2008.
[16] L. H. Nielsen, E. Løkkegaard, A. H. Andreasen, Y. A. Hundrup, and N. Keiding, "Estimating the effect of current, previous and never use of drugs in studies based on prescription registries," *Pharmacoepidemiol. Drug Safety*, vol. 18, no. 2, pp. 147–153, 2009.
[17] M. Toussi, J.-B. Lamy, P. Le Toumelin, and A. 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 Informat. Decis. Making*, vol. 9, no. 1, p. 28, 2009.
[18] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. IEEE 11th Int. Conf. Data Eng.*, 1995, pp. 3–14.
[19] E. Lalla and P. N. Papapanou, "Diabetes mellitus and periodontitis: A tale of two common interrelated diseases," *Nature Rev. Endocrinol.*, vol. 7, no. 12, pp. 738–748, 2011.
[20] D. Bagley, "Parallel protocols: Treating diabetes and HIV/AIDS," November 2015, [Online]. Available: https://endocrinenews.endocrine.org/parallel-protocols-treating-diabete s-and-hivaids/