T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title	Sequential Pattern Mining on Electronic Medical Records with Handling Time Intervals and the Efficacy of Medicines
Author	Keishiro Uragaki, Tomoyuki Hosaka, Yoshitaka Arahori, Muneo Kushima, Tomoyoshi Yamazaki, Kenji Araki, Haruo Yokota
Journal/Book name	Proc. of the 21st IEEE International Symposium on Computers and Communications, , ,
Issue date	2016, 6
DOI	http://dx.doi.org/10.1109/ISCC.2016.7543708
URL	http://www.ieee.org/index.html
Copyright	(c)2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.
Note	このファイルは著者（最終）版です。 This file is author (final) version.

Powered by T2R2 (Science Tokyo Research Repository)

Sequential Pattern Mining on Electronic Medical Records with Handling Time Intervals and the Efficacy of Medicines

Keishiro Uragaki*, Tomoyuki Hosaka*, Yoshitaka Arahori*, Muneo Kushima**,
Tomoyoshi Yamazaki**, Kenji Araki** and Haruo Yokota*
*Department of Computer Science, Tokyo Institute of Technology,
2-12-1 Oookayama, Meguro-ku, Tokyo 152, Japan
Email:yokota@cs.titech.ac.jp
**Faculty of Medicine, University of Miyazaki Hospital,
5200 Machikihara, Kiyotake-Cho, Miyazaki-shi, Miyazaki 889, Japan
Email: jyoho_support@med.miyazaki-u.ac.jp

***Abstract*—It is useful to employ electronic medical records to improve medical studies. Based on their experience, medical workers conventionally prepare *clinical pathways* as guidelines for the typical flow for the medical treatment of each disease. In this study, we propose an approach for verifying existing clinical pathways and recommend variants or new pathways by analyzing historical records. We propose a method based on the application of sequential pattern mining to record logs with handling time intervals between treatments. We also focus on the efficacy of medicines instead of their names because various medicines have the same efficacy and they change dynamically. We evaluated the proposed method using actual logs and the results demonstrated that the proposed method is effective.**

I. Introduction

Most large-scale hospitals have recently adopted electronic medical record (EMR) systems to support medical workers with the simple maintenance and checking of patient information, which can be compared rapidly with written records. Moreover, the secondary uses of electronic medical records have attracted attention with respect to the standardization of medical treatments. Medical workers tend to use *clinical pathways* as guidelines for their medical actions. A clinical pathway defines the typical flow for the medical treatment of each disease and it is conventionally generated by medical workers based on their experience. However, this human-based approach is very time-consuming and much effort is required to produce correct pathways, while there is also the possibility of including inappropriate components in the results obtained. Clinical pathways can be extracted from long historical EMRs, which are kept in hospitals for secondary usage. Machine-generated clinical pathways can be used to verify existing clinical pathways and to recommend variants or new pathways.

In this study, we propose a method based on the application of a sequential pattern mining algorithm to EMR logs for extracting clinical pathways. The time intervals between medical treatments are important, such as between checkups and doses; therefore, the proposed method considers the time intervals between medical treatments. Moreover, various medicines have the same efficacy and new medicines may appear dynamically, so we consider the efficacy of medicines instead of their names in the proposed method. We used actual EMR logs from a university hospital to evaluate the proposed method and the experimental results demonstrated that the proposed method is effective.

The remainder of this paper is organized as follows. Related research is reviewed in Section II. The proposed method is described and evaluated in Sections III and IV, respectively. We provide our conclusions in Section V.

II. Related Work

Many studies have addressed the secondary uses of EMRs related to clinical pathways. Okada et al. proposed an electronic clinical pathway system based on a semistructured data model that uses a personal digital assistant for on-site access [1]. Wakamiya and Yamauchi considered standard functions for electronic clinical pathways [2]. Hirano and Tsumoto proposed a method for extracting typical medical processes by analyzing the logs stored in a hospital information system [3], where they clustered the medical treatment data from each implementation date to create candidate clinical pathways, although they did not employ a sequential pattern mining approach.

In this study, we propose the application of a sequential pattern mining algorithm to EMRs to extract clinical pathways. The sequential pattern mining algorithm is an Apriori-based frequent pattern mining algorithm, which is well known [4]. However, this algorithm is very time-consuming with large data sets and it generates a large number of irrelevant patterns as results. Moreover, it cannot handle time intervals between items.

The *time-interval sequential pattern mining (TI-SPM)* algorithm was proposed to handle the time intervals between items [6]. Because the TI-SPM algorithm is based on PrefixSpan [7], it is much faster than the Apriori-based algorithm. However, TI-SPM still generates irrelevant patterns, so the concept of *closed sequential pattern mining* [5] can be applied to remove irrelevant patterns from the results.

TI-SPM represents the time interval between items using a *time interval* set, which is a group that comprises a finite number of numeric values. *Multi-TI-SPM (MTI-SPM)* was proposed [8] for determining the time interval between multiple items in patterns. Both TI-SPM and MTI-SPM are effective when the time interval between items does not change greatly, but they are not suitable if the time interval between items changes in a dynamic manner, as found with medical treatments.

In this study, we propose a method called *T-PrefixSpan* for calculating the time interval between medical treatments based on statistical information to extract clinical pathways. Although Huang et al. employed a similar approach [9], ours differs with respect to the calculation methods and medicinal treatments, especially the efficacy of medicines. Lin et al. proposed a method for mining time dependency patterns in clinical pathways [18], which is another approach to the secondary usage of EMRs that considers time.

III. Methods

A. Handling Time Interval between Items

In medical care, it is essential to manage the time interval during sequential medical treatments. The existing TI-SPM [6] handles a

set of time intervals by dividing them into time ranges. However, more accurate information about time is required to specify clinical pathways. Thus, we propose a method called *T-PrefixSpan* for deriving statistical information about time intervals by considering outliers. Huang et al. proposed a similar approach for calculating time intervals based on sequential pattern mining [9], but their approach to the statistical calculations differs from ours because they did not handle outliers or consider the efficacy of medicines.

First, we define the concepts required to introduce our method before we explain T-PrefixSpan.

Definition 1: **T-item** (i, t)
Let I be the set of items and t is the time when an item i has occurred. We define a **T-item** (i, t) as a pair of i and t.

Definition 2: **T-sequence** s and **O-sequence** O_s
T-sequence s is a sequence of T-items, which is denoted by $s =< (i_1, t_1), (i_2, t_2), ..., (i_n, t_n) >$. T-items that occur at the same time should be arranged in alphabetical order. Furthermore, let n be the length of T-sequence s and let an **O-sequence** of s be the sequence $O_s =< i_1, i_2, ..., i_n >$.

Definition 3: **time interval** TI_k
Given a T-sequence $s =< (i_1, t_1), (i_2, t_2), ..., (i_n, t_n) >$, **time interval** TI_k is defined as follows:

$$TI_k \equiv t_{k+1} - t_k \ \ (k = 1, 2, ..., n-2, n-1).$$

Definition 4: **T-sequential database** D and **O-sequential database** O_D
Given a set of T-sequences, **T-sequential database** D is defined as follows, where the identifier sid of the element of D has a unique value for each sequence.

$$D \equiv \{(sid, s) \mid sid, s \in S\}$$

Furthermore, let an **O-sequential database** O_D be a sequential database that comprises O-sequences configured from all of the T-sequences in D. Let $Size(D)$ be $Size(O_(D))$, which is the number of sequences in O_D

Definition 5: **T-frequent sequential pattern** P
Let $MinSup$ $(0 \leq MinSup \leq 1)$ be a minimum support and D is a T-sequential database. Given $P =< i_1, X_1, i_2, X_2, ..., i_{n-1}, X_{n-1}, i_n >$ ($\forall j$ i_j is an item, $\forall k$ X_k is a set of five values, i.e., $(min_k, mod_k, ave_k, med_k, max_k)$, and we can configure a sequence $O_P =< i_1, i_2, ..., i_{n-1}, i_n >$.
We define P as a **T-frequent sequential pattern** if O_p is a frequent sequential pattern in an O-sequential database configured from D (i.e., $Sup(P) =\mid \{Seq \mid O_p \subseteq Seq, (sid, Seq) \in O_D$, where sid is an identifier of $Seq\} \mid \geq Size(O_D) \times MinSup$).
Let O_P be the O-Pattern of P. The set of five values is defined as follows.
Given all of the T-sequences with O-sequences containing O_P in D, let S be one of them, where $S =< i'_1, t_1, i'_2, t_2, ..., i'_{m-1}, t_{m-1}, i'_m >$. By using $j_1, j_2, ..., j_{n-1}, j_n$, which satisfies (1) $1 \leq j_1 < j_2 < ... < j_{n-1} < j_n \leq m$ and (2) $i_k = i'_{j_k}, i_{k+1} = i'_{j_{k+1}}$, we can configure sets of time intervals: $Set_{TI_1}, Set_{TI_2}, ..., Set_{TI_{n-1}}$, where $TI_k = t'_{j_{k+1}} - t'_{j_k}$. Moreover, in $X_k = (min_k, mod_k, ave_k, med_k, max_k)$, we define the five values as follows.

1) $min_k = \min Set_{TI_k}$
2) mod_k = the most frequent value in Set_{TI_k}
3) ave_k = the average of the values in Set_{TI_k}
4) med_k = the intermediate value of the values in Set_{TI_k}
5) $max_k = \max Set_{TI_k}$

Given a time interval $X_j = (min_j, mod_j, ave_j, med_j, max_j)$ $(1 \leq j < n)$, if the equation $min_j = Max_j$ holds, then the time interval between item i_j and item i_{j+1} is consistent; in particular, if the equation $min_j = Max_j = 0$ holds, then these two items occurred at the same time.

Definition 6: **T-closed frequent sequential pattern** A
Given a T-sequential database D, let $\sum$ be a set of T-frequent sequential patterns extracted from D and A is a T-frequent sequential pattern of $\sum$. A is a **T-closed frequent sequential pattern** if B satisfying the following does not exist in $\sum \backslash A$.

1) If we let A' and B' be O-Patterns of A and B, respectively, then $A' \subseteq B'$.
2) $Sup(A) \leq Sup(B)$, where we define a support of the T-frequent sequential pattern as $Sup(A) \equiv\mid \{s \mid s \subseteq S, (sid, S) \in D, wheresid$ is the identifier of S in $D\} \mid$.
3) If we let A and B be $< a_1, T_1, a_2, T_2, ..., a_{n-1}, T_{n-1}, a_n >$ and $< b_1, T'_1, b_2, T'_2, ..., b_{m-1}, T'_{m-1}, b_m >$, respectively, the $j_1, j_2, ..., j_n$ exist that satisfy (1) $1 \leq j_1 < j_2 < ... < j_n \leq m$ and (2) $a_k = b_{j_k}, a_{k+1} = b_{j_{k+1}}$. Thus, for all $T_k = (min_k, mod_k, ave_k, med_k, max_k)$ and $T'_{j_k} = (min'_{j_k}, mod'_{j_k}, ave'_{j_k}, med'_{j_k}, max'_{j_k})$, equations (1) $min_k \geq min'_{j_k}$ and (2) $max_k \leq max'_{j_k}$ hold.

For example, if we are extracting T-frequent sequential patterns from a T-sequential database such as Table I under the minimum support $MinSup = 0.4$, then because the O-sequential database of D is shown in Table II, the frequent sequential patterns based on the minimum support $MinSup = 0.4$ are $< A >$, $< B >$, $< E >$, $< A, B >$, $< B, E >$, and $< A, B, E >$. The frequent sequential patterns with one item in O_D are T-frequent sequential patterns in D, so $< A >$, $< B >$, and $< E >$ are T-frequent sequential patterns in D.

Considering the time between item A and item B in the sequence $< A, B >$, the set of time intervals calculated from D is $\{2, 3, 3, 4, 5\}$. By considering the minimum, most frequent value, average, median, and maximum, then $< A, (2, 3, 3, 3, 5), B >$ is a T-frequent sequential pattern in D (2+3+3+4+5=17, [17/5]=3). Similarly, if we calculate the T-frequent sequential pattern from $< B, E >$ and $< A, B, E >$, we have $< B, (3, 5, 5, 5, 7), E >$ and $< A, (2, 2, 2, 2, 3), B, (3, 5, 5, 5, 7), E >$. If more than two different values are the most frequent values, then their average is the most frequent value. Therefore, the T-frequent sequential patterns in D under the minimum support $MinSup = 0.4$ are $< A >$, $< B >$, $< E >$, $< A, (2, 3, 3, 3, 5), B >$, $< B, (3, 5, 5, 5, 7), E >$, and $< A, (2, 2, 2, 2, 3), B, (3, 5, 5, 5, 7), E >$, and the T-closed frequent sequential patterns are $< A >$,$< B >$, $< A, (2, 3, 3, 3, 5), B >$ and $< A, (2, 2, 2, 2, 3), B, (3, 5, 5, 5, 7), E >$.

We developed T-PrefixSpan from PrefixSpan [7]. The T-PrefixSpan algorithm is described in **Algorithm 1**, where an O-sequential database of a T-sequential database D is O(D), an O-sequence of a T-sequence S is O(S), and the connection of sequence A with sequence B is AB. Furthermore, the N-th element of a set X is X_N, the N-th item of a sequence S is S_N, and the time when the N-th T-item of T-sequence A occurred is T_{A_N}.

We use the Smirnov–Grubbs test as an arbitrary function to exclude outliers from the set of time intervals, where we exclude outliers below the significance level of $\alpha = 0.05$.

TABLE I
T-SEQUENTIAL DATABASE D

Identifier for a T-sequence	T-sequence
10	$< (A,1),(B,3),(C,7),(E,10) >$
20	$< (A,1),(B,4),(E,7) >$
30	$< (A,2),(B,6),(B,9) >$
40	$< (A,2),(B,5) >$
50	$< (A,2),(B,7) >$

TABLE II
O_D (O-SEQUENTIAL DATABASE OF D)

Identifier for a T-sequence	T-sequence
10	$< A,B,C,E >$
20	$< A,B,E >$
30	$< A,B,B >$
40	$< A,B >$
50	$< A,B >$

B. Handling Medicines

We represent a medical treatment item as a set of four subitems: (*Class*; *Description*; *Code*; *Name*). *Class* denotes the classification of a medical treatment, *Description* is its detailed diagnostic record, *Code* is a medicinal code that represents the unique efficacy of the medicine considered, and *Name* is the name of the medicine. The total number of *Class* types is finite. If a medical treatment administered to a patient is not a medicine, *Code* and *Name* are set to "null" in order to clarify that they are blank.

For example, when a medical treatment designated as "prescribe" an "internal medicine" where the medicinal Code is "613" and the name is "Fine Cefzon 10% for Children" appears in a log, the item is represented in the form: (prescription; internal medicine; 613; Fine Cefzon 10% for Children). In this example, *Code* 613 is an "antibiotic preparation that acts mainly against Gram-positive, Gram-negative bacteria." If we need to represent a medical treatment that comprises a "nursing task" of "replacing the sheets," then the item is represented in the form: (nursing task; replacing the sheets; null; null).

Next, we introduce some techniques for organizing a medical treatment as an item. First, the medicines used during surgery vary for each patient, so the *Code* and *Name* of the medicine employed in the medical treatment where the *Class* does not include a word related to surgery (e.g., "surgery" and "anesthesia") are set as "null" because the information about the medicine is not medically useful. Second, if the medical treatment is determined only by the contents of the *Class* such as "sampling" (which depends on the EMR), then the *Description*, *Code*, and *Name* are set as "null" because they are not medically useful. Next, depending on the language used in an EMR, *Description* is the first (or last) clause of a sentence in the description given by a medical worker because the important information is written in the first (or last) clause. In Japanese, the first clause is important. Finally, we delete the sequences that do not contain items where the *Class* lacks words related to surgery because they are not medically useful. We specify the group of four medical data values (Class, Description, Code, Name) as an item and we configure a T-item by adding the time when a medical treatment is applied. We configure a T-sequence comprising T-items, which are medical treatments received by one patient until their discharge from hospital. If the T-items occurred at the same time in a T-sequence and the items in the T-items are the same, we regard these T-items as one item. We sort the T-items in a T-sequence alphabetically according to the order of Class, Description, Code, and Name. Finally we configure a T-sequential database based on these T-sequences. We consider T-sequences as different T-sequences if the patients are the same but their admission–discharge periods differ.

Algorithm 1 T-PrefixSpan

Input: D: a T-sequential database, $MinSup$: a minimum support
Output: P: the set of T-frequent sequential patterns
Call: T-PrefixSpan($<>$,D)
Procedure: T-PrefixSpan(α,$D\mid_\alpha$)

```
D' |_α = O(D |_α)
if α != null then
    P ← GetProperTime(α, D |_α, D' |_α)
end if
B ← {β | (s ⊆ D' |_α, β ∈ s) ∧ (Sup(β) ≥ Size(D) × Minsup)}
for β ∈ B do
    D |_αβ ← {< sid, s > ∈ D |_α | αβ ⊆ O(s)}
    Call T-PrefixSpan(αβ, D |_αβ)
end for
```

Subroutine: GetProperTime(α,$D\mid_\alpha$,$D'\mid_\alpha$)

```
if length(α) == 1 then
    return α
end if
K ← {k |< sid, s > ∈ D |_α, O(s) ∈ D' |_α, k ⊆ s, O(k) == α}
T = {{}, {}, ..., {}}(| T |= length(α) − 1)
for k ∈ K do
    for i = 0, ..., length(k − 1) do
        T_i ← T(k_{i+1}) − T(k_i)
    end for
end for
W =< α_0, α_1, ..., α_{length(α)−1} >
for i = 0, ..., length(α) − 2 do
    T_i = an arbitrary function for excluding outliers from T_i
    min_i = min T_i
    mod_i =the most frequent value of T_i
    ave_i =the average of the values of T_i
    med_i =the intermediate value of the values of T_i
    max_i = max T_i
    X_i = (min_i, mod_i, ave_i, med_i, max_i)
    W =< α_0, ..., α_i, X_i, α_{i+1}..., α_{length(α)−1} >
end for
return W
```

C. Efficacy of Medicines

In the previous section, we described the representation of an item as a set of four text types because it is difficult to extract patterns with only the *Class*, *Description*, and *Name* fields as various medicines have the same efficacy and they may change dynamically. Wright et al. [10] conducted mining by focusing on attributes other than the name of the medicine, where they extracted frequent patterns from a sequential database based on sequences that contained items represented by medicines administered to diabetic patients. Their method focused on the classification of medicines and they regarded items as the same if their classifications matched with each other. The results obtained showed that their method focusing on the classification of medicines could predict the next item for administration with a higher probability compared with a naive method that lacked this focus.

Based on the method to focus on attributes other than names of medicines, our proposed method focuses on the efficacy of medicines and we regard items as the same if their *Class*, *Description*, and *Code* match even if their *Name* differs, so we can extract medically useful

patterns by reducing the number of item types. A related method that focused on the efficacy of medicines was described by [11].

IV. Experiment

Based on our proposed method for handling information about medicines and the time intervals between items, we evaluated the changes in the output when we focused only on the efficacy of medicines compared with the actual data.

A. Experimental Data

We used target medical treatment data based on the clinical pathways recorded from November 19, 1991 to October 4, 2015 in the EMRs at the Faculty of Medicine, University of Miyazaki Hospital. These medical data were acquired using an EMR system *WATATUMI* [12] employed by the Faculty of Medicine, University of Miyazaki Hospital. We did not include information that could identify a patient uniquely to ensure the protection of personal information.

When we extracted the medical treatment data, we used anonymous patient IDs, which were impossible to determine. The data we extracted from the EMRs to support medical treatments at the Faculty of Medicine, University of Miyazaki Hospital were described previously in [13] and they can be accessed at the website of the University of Miyazaki. Our study was approved by the Ethics Review Board of the University of Miyazaki and the Research Ethics Review Committee of Tokyo Institute of Technology.

Our target data comprised medical treatments based on two clinical pathways: (1) *cryptorchidism fusion surgery* and (2) *TUR-Bt (Transurethral Resection of the Bladder tumor)*, which were included in the EMRs. We selected these two clinical pathways because (1) *cryptorchidism fusion surgery* is a clinical pathway where the flow of the medical treatments was fixed, whereas (2) *TUR-Bt* is a clinical pathway for which the flow was not clearly defined.

B. Experimental Procedure

In the experiment, we compared two methods for using the minimum support, as follows: (1) a "*Normal*" method that extracted patterns using the four values, i.e., *Class*, *Description*, *Code*, and *Name*; and (2) a "*Focusing on Efficacy*" method that extracted patterns using only three values, i.e., *Class*, *Description*, and *Code*. The output patterns were extracted by T-PrefixSpan. The indicators compared were the number of outputs, the average length of the outputs, and the proportion of patterns including medical treatments that involved medicines relative to the total. After the comparative experiments, we checked whether the typical flow obtained using our extraction method agreed with the clinical pathway prepared by a medical doctor based on their experience.

The number of sequences, average length of the sequences, minimum length of the sequences, and maximum length of the sequences for the two data sets, i.e., (1) *cryptorchidism fusion surgery* and (2) *TUR-Bt*, are shown in Table III.

TABLE III
Target Dataset

dataset	*Cryptorchidism fusion surgery*		*TUR-Bt*	
	Normal	*Focusing on Efficacy*	*Normal*	*Focusing on Efficacy*
Number of sequences	265		488	
Average length	19.64	19.16	53.21	49.89
Minimum length	10	9	11	11
Maximum length	460	465	655	485

C. Results and Discussion

The numbers of outputs are shown in Fig. 1 and Fig. 2, the average lengths of the outputs are shown in Fig. 3 and Fig. 4, and Fig. 5 and Fig. 6 show the proportion of patterns including medical treatments that involved medicines relative to the total.

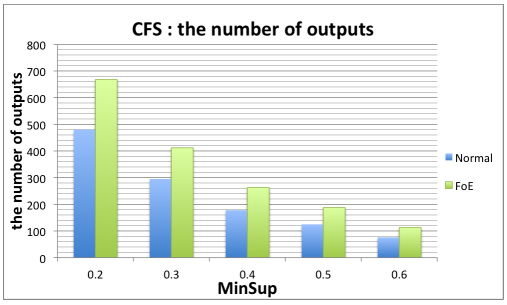


Fig. 1. *Cryptorchidism fusion surgery*: number of outputs

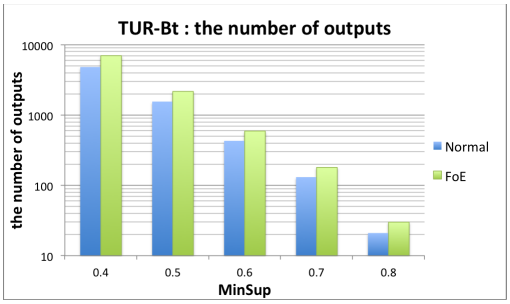


Fig. 2. *TUR-Bt*: number of outputs

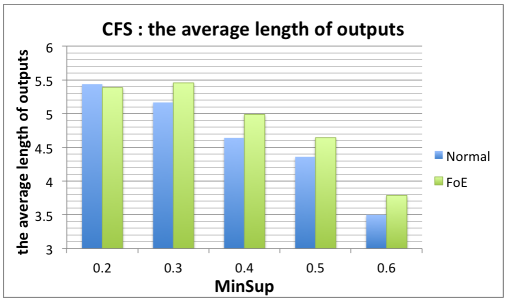


Fig. 3. *Cryptorchidism fusion surgery*: average length of outputs

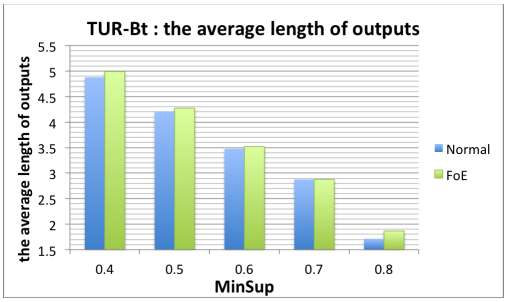


Fig. 4. *TUR-Bt*: average length of outputs

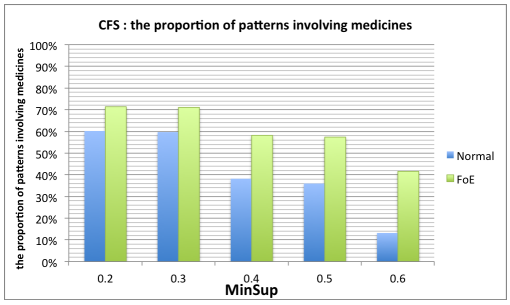


Fig. 5. *Cryptorchidism fusion surgery*: proportion of patterns involving medicines

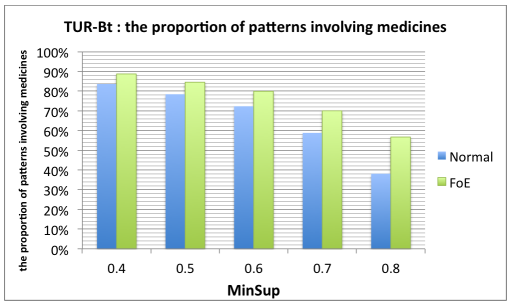


Fig. 6. *TUR-Bt*: proportion of patterns involving medicines

Based on the results of the experimental comparisons, we determined the following. Figures 1 and 2 indicate that the number of sequences was higher using the *Focusing on Efficacy* method compared with that using the *Normal* method with either *cryptorchidism fusion surgery* or *TUR-Bt*. There was an increase in the variety of items for which the frequency exceeded the minimum support because we regarded items with different *Name* fields as the same item if they had the same efficacy.

The execution time was higher using the *Focusing on Efficacy* method compared with that using the *Normal* method because the recursion number increased with T-PrefixSpan when the frequency of the number of item types exceeded the minimum support. According to Fig. 3 and Fig. 4, the average length obtained using the *Focusing on Efficacy* method tended to be larger than that using the *Normal* method for both data sets, because medical workers frequently administer medicines with the same name that differ in their efficacy among patients. As shown in Fig. 5 and Fig. 6, the proportion of patterns involving medicines was higher using the *Focusing on Efficacy* method compared with that using the *Normal* method, which shows that the former method could extract more medical treatments that involved medicines than the *Normal* method.

For all of the outputs obtained under minimum support values ranging from 0.02 to 1.0, with surgery as day 0 and the items that occurred on each implementation date, we found that the flow included most of the outputs, as shown in Fig. 7. It is not possible to extract latent medical treatments by normal data mining according to [14]. Thus, we regarded items with the same probability of more than 50% of occurring the day before the surgery as one item. Using this method, we extracted the items that appeared only with low support.

In Fig. 7, a blue item represents surgery, red items are the medical treatments that involved medicines, and green items are medical treatments that did not involve medicines. Figure 7 shows a common set of patterns, including surgery, where the time interval between two items was constant when their minimum and maximum values matched. We excluded medical treatments with indefinite

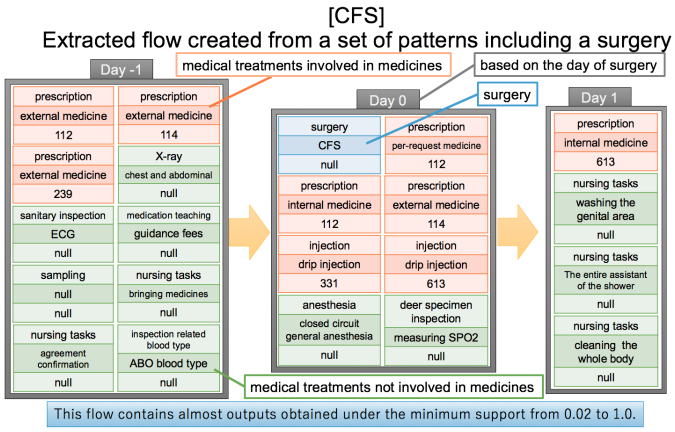


Fig. 7. *Cryptorchidism fusion surgery*: the typical flow obtained by our extraction method

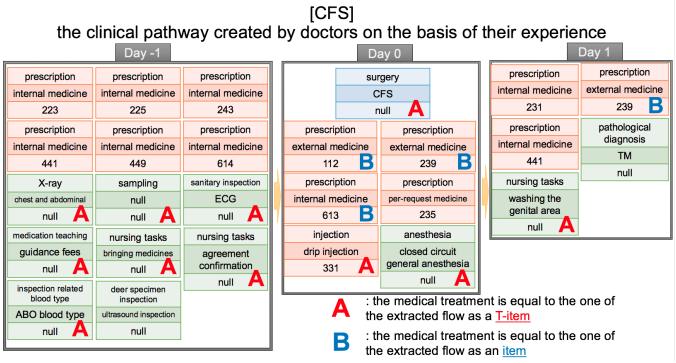


Fig. 8. *Cryptorchidism fusion surgery*: the clinical pathway produced by a medical doctor based on their experience

Medicinal classification table	
Code	Efficacy that corresponds to Code
112	Hypnotics and sedatives, antianxietics
114	Antipyretics, analgesics and antiinflammatory agents
223	Expectorants
225	Bronchodilating preparations
231	Antidiarrheals, intestinal regulators
235	Purgatives and clysters
239	Other agents affecting digestive organs
243	Thyroid and para-thyroid hormone preparations
331	Blood substitutes
441	Antihistaminic
449	Other antiallergic agents
613	Antibiotic preparations acting mainly on gram-positive and gram-negative bacteria
614	Antibiotic preparations acting mainly on gram-positive bacteria and mycoplasma

Fig. 9. *Cryptorchidism fusion surgery*: medicinal classification table

implementation dates. This approach loosely represents items that occur on the same day as a daily set, so it is possible that a branch exists.

A medicinal classification table representing the *Code* field and efficacy is shown in Fig. 9, which was created based on [15], as provided by the Ministry of Health, Labor, and Welfare in Japan. Figure 8 shows the clinical pathway created by medical doctors according to their experience, where medical treatments with indefinite implementation dates were excluded. In our comparison of Fig. 7 and Fig. 8, items with consistent implementation dates are denoted by red "A" and the items with implementation dates that differ from those in the clinical pathway are denoted by blue "B". Thus, by comparing Fig. 7 and Fig. 8, we can see that for the medical treatments involving medicines, a high proportion of the item contents and implementation dates were consistent, whereas for the medical treatments that did not involve medicines, many of the implementation dates in the outputs did not match with those in the clinical pathway.

Thus, the actual flows of the medical treatments differed from the flows assumed by the medical workers. We could not extract some items involving medicines from the datasets because we could not obtain the items by mining under the minimum support value of $MinSup = 0.02$. It will be necessary to determine whether the processes that did not match with the actual pathway are medically important based on discussions with medical workers.

V. Conclusions

In this study, we proposed a method for extracting clinical pathways from historical EMRs. We developed T-PrefixSpan as a sequential pattern mining algorithm, which can handle time intervals to derive their statistical information, as well as considering the efficacy of medicines instead of their names. In the proposed method, medical workers can refer to the time intervals between medical treatments based on five values, i.e., the minimum, mode, average, median, and maximum.

We evaluated the proposed method using actual logs from a university hospital. The experimental results showed that the number of patterns obtained by focusing on the efficacy of medicines was greater than that based on the names of medicines. We obtained a typical flow for cryptorchidism fusion surgery and the result was similar to the clinical pathway generated by medical doctors based on their experience regarding items not involving medicines. The generated patterns can be used to verify existing clinical pathways and to recommend variants or new pathways. In future research, we plan to evaluate the proposed method with other diseases. In addition, to extend the proposed method, it will be necessary to introduce a faster algorithm for mining under a low minimum support, which cannot be calculated in real time with T-PrefixSpan. We aim to consider the suitability of faster sequential pattern mining algorithms such as CloSpan [5], Clasp [16], and CSpan [17] for time-interval-based extraction.

The method used to measure the similarity between the medical treatments applied to patients and the clinical pathway was described in [19]. During evaluations, we would normally expect to the use the minimum and maximum values of the time intervals between items, but it is possible to incorporate the most frequent value, average, and median.

Medical workers should be asked to evaluate whether the outputs are medically important. It would be useful to develop a sophisticated technique for presenting the outputs to medical workers by considering the branching flow of medical treatments.

Acknowledgements

This research was supported in part by a JSPS Grant-in-Aid for Scientific Research (A) (#25240014).

References

[1] Osamu Okada, Naoki Ohboshi, Tomohiro Kuroda, Keisuke Nagase, and Hiroyuki Yoshihara. Electronic clinical path system based on semistructured data model using personal digital assistant for onsite access. Journal of Medical Systems 29 (4), 379–389, 2005.

[2] Shunji Wakamiya and Kazunobu Yamauchi. What are the standard functions of electronic clinical pathways? International Journal of Medical Informatics 78, 543–550, 2009.

[3] Shoji Hirano and Shusaku Tsumoto. Clustering of order sequences based on the typicalness index for finding clinical pathway candidates. The IEEE International Conference on Data Mining (ICDM Workshops), 2013.

[4] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, 487–499, 1994.

[5] Xifeng Yan, Jiawei Han and Ramin Afshar. CloSpan: Mining closed sequential patterns in large databases. Proceedings of the 2003 SIAM International Conference on Data Mining, 166–177, May 2003.

[6] Yen-Liang Chen, Mei-Ching Chiang and Ming-Tat Ko. Discovering time-interval sequential patterns in sequence databases. Expert Systems with Applications 25, 343–354, 2003.

[7] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal and Mei-Chun Hsu. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. Proceedings of 2001 International Conference on Data Engineering, 215–224, 2001.

[8] Ya-Han Hu, Tony Cheng-Kui Huang, Hui-Ru Yang and Yen-Liang Chen. On mining multi-time-interval sequential patterns. Data & Knowledge Engineering 68, 1112–1127, 2009.

[9] Zhengxing Huang, Xudong Lu and Huilong Duan. On mining clinical pathway patterns from medical behaviors. Artificial Intelligence in Medicine 56, 35–50, 2012.

[10] 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.

[11] Sang-Jun Yea, BoSeok Seong, Yunji Jang and Chul Kim. A data mining approach to selecting herbs with similar efficacy: Targeted selection methods based on medical subject headings (MeSH). Journal of Ethnopharmacology 182, 27–34, 2016.

[12] *Denshi Karte System WATATUMI* (EMRs "*WATATUMI*"). http://www.corecreate.com/02_01_izanami.html

[13] *Miyazaki Daigaku Igaku Bu Fuzoku Byouin Iryo Jyoho Bu* (Medical Informatics Division, Faculty of Medicine, University of Miyazaki Hospital). http://www.med.miyazaki-u.ac.jp/home/jyoho/

[14] Zhengxing Huang, Wei Dong, Lei Ji, Chunhua He and Huilong Duan. Incorporating comorbidities into latent treatment pattern mining for clinical pathways. Journal of Biomedical Informatics 59, 227-239, 2016.

[15] KEGG MEDIUS Drug Classification http://www.genome.jp/kegg/medicus/drugclass.html

[16] Antonio Gomariz, Manuel Campos, Roque Marin and Bart Goethals. Clasp: An efficient algorithm for mining frequent closed sequences. Advances in Knowledge Discovery and Data Mining 7818, 50-61, 2013.

[17] V. Purushothama Raju and G.P. Saradhi Varma. Mining closed sequential patterns in large sequence databases. International Journal of Database Management Systems 7.1, 29-39, 2015.

[18] Fu-Ren Lin, Shien-Chao Chou, Shung-Mei Pan and Yao-Mei Chen. Mining time dependency patterns in clinical pathways. International Journal of Medical Informatics 62.1, 11–25, 2001.

[19] Zhengxing Huang, Xudong Lu and Huilong Duan. Similarity measuring between patient traces for clinical pathway analysis. Artificial Intelligence in Medicine 7885, 268–272, 2013.