Analysis of Transitions in Differences between Frequent Medical-order Sequences for COVID-19

Zitai Zhao
Tokyo Institute of Technology
School of Computing
Tokyo, Japan
kotaidesu@gmail.com

Yuki Yasumitsu
Tokyo Institute of Technology
School of Computing
Tokyo, Japan
yasumitsu@de.cs.titech.ac.jp

Hieu Hanh Le
Tokyo Institute of Technology
School of Computing
Tokyo, Japan
hanhlh@de.cs.titech.ac.jp

Ryosuke Matsuo
Life Data Initiative
Tokyo, Japan
matsuo@ldi.or.jp

Tomoyoshi Yamazaki
Tokyo Institute of Technology
School of Computing
Tokyo, Japan
yamazaki.cp@gmail.com

Kenji Araki
Faculty of Medicine
University of Miyazaki Hospital
Tokyo, Japan
araki6925@gmail.com

Haruo Yokota
Tokyo Institute of Technology
School of Computing
Tokyo, Japan
yokota@cs.titech.ac.jp

***Abstract*—With the increasing use of electronic medical records, medical support from analysis of the accumulated medical information is expected. Currently, new treatment methods and drugs are being developed for the treatment of new diseases, but the transition history of medical orders has yet to be visualized for diseases such as COVID-19.**

In this paper, we use sequential pattern mining to extract frequent medical orders and then apply the longest common subsequence variant (LCSV) and merged sequence variant (MSV) to analyze the differences in treatment patterns at different times. We also propose three types of sliding window (time interval window, sequence number window, and time-sequence number window) to analyze the transition history of medical orders. As an example, we applied these methods to Japanese electronic medical records covering the first to the fifth waves of COVID-19 and analyzed the differences in medical-order patterns for the five infection waves and the transition history of medical orders. We then visualized the difference with MSV. The results showed that the proposed method can successfully visualize the differences in medical orders between infection waves, and the transition history of medical orders can be revealed. The validity of the results was confirmed by the medical staff involved.

***Index Terms*—Electronic health records, sequential pattern mining, sequence variants, COVID-19**

I. Introduction

In recent years, the use of electronic medical records (EMRs) has been increasing, and this trend is expected to continue in the future. This has led to the expectation of secondary use of the accumulated medical information. For example, there have been studies using sequential pattern mining (SPM) to extract frequent treatment patterns for the analysis of EMR for medication recommendations [1]. In the past, clinical pathways were created based on the medical experience of medical staff, which was not easy to collect and analyze using human resources. Therefore, research is being conducted to support medical practice improvement by analyzing EMRs from a data engineering perspective. In a new development, the spread of COVID-19 has led to the emergence of mutant strains and the development of new treatment methods and drugs, which has resulted in changes in medical-order patterns depending on the time period. Understanding the differences in medical-order patterns across infection waves and the transition history of medical orders may well be useful in future infectious disease control and may contribute to improving medical care overall.

This paper aims to help provide such medical support by analyzing the differences in medical-order sequences for a disease during different periods and the transition history of medical orders to assess the impact of variant strains, treatments, and drugs on medical-order sequences.

The effectiveness of our proposed method was evaluated using the EMR data of patients with COVID-19 who were hospitalized and treated. The results showed that the proposed method can successfully visualize the differences in medical orders between infection waves and the transition history of medical orders. This result was confirmed independently via the cooperation of medical staff.

Note that the proposed method is applicable not only to COVID-19 but also to other diseases.

The contributions of this paper are as follows.

1) In this paper, merged subsequence variant (MSV), and the longest common subsequence variant (LCSV) [4], previously used in a unique approach proposed by our research group to show differences in medical orders for the same procedure across providers, are used to analyze differences in medical orders for different infection waves of COVID-19.
2) We propose three types of sliding window (time interval window, sequence number window, and time-sequence number window) to obtain the temporal characteristics of patterns by adjusting the window size and slide unit for each window. We can then analyze the transition history of medical orders by evaluating the similarity between adjacent windows.

3) By experimenting with the actual dataset for COVID-19 in the EMR systems of two medical institutions, the differences in treatment patterns between different waves of infection and between different medical institutions could be visualized. Using the proposed sliding windows, the transition of medical orders could also be visualized. The effectiveness of these results was confirmed by medical staff.

The remainder of this paper is organized as follows. Related work is summarized in Section II. The proposed methods and experimental evaluation are described in Sections III, IV, respectively. Conclusions are discussed in Section V.

II. Related Work

A. SPM

SPM, proposed by Agrawal et al. [1], is a method for extracting frequent patterns from a sequence database (SDB) and has attracted attention in domains such as medicine, e-commerce, and Internet-related studies. A sequence of items is called a "sequence", and an SDB comprises elements that are composed of sequences belonging to a certain set of sequences and sequence identifiers.

B. Analysis of medical data via SPM

SPM has been widely used in studies of medical information, particularly in medical-order sequences. Le et al. proposed T-CSPan [2], which is an extension of CSpan, an efficient SPM, to collect time-interval statistics and applied it to actual EMR data. On the next-item recommendation for large combinations of items with varied values, Le et al. proposed an effective method to recommend the next item set given the query item set and the associated values [5].

C. Analysis of SVs

In the frequent medical-order patterns extracted by SPM, there may be variants (i.e., SVs) in which just some parts of the pattern differ from other patterns. Such SVs can correspond to the medical care that varies depending on the patient ' s condition. An SV is generated when a split occurs in a medical-order sequence. Honda et al. [3] extracted SVs in medical orders from real EMRs and presented them as graphs in a visualization tool for medical staff.

D. Comparison of SVs from multiple medical institutions

Whereas previous analyses have involved single medical orders, Li et al. [4] extracted the LCSV of two SVs to show the differences in frequently occurring medical-order sequences containing different SVs in different medical orders and proposed the concept and a derivation algorithm for the MSV. Then, by applying these concepts to the EMR data from two medical institutions, the differences in treatment patterns between the two institutions became apparent.

The use of the LCSV and MSV algorithms can establish the MSV of two SVs. The MSV shows the common part of the two SVs and where the branches appear, by labeling. This makes it easy for medical workers to recognize the differences in treatment patterns and thereby improve the clinical pathways.

III. Method

A. Appropriate data preprocessing based on information from the medical staff

Preprocessing enables exclusion of the following inappropriate medical orders.

- Medical orders that are medically irrelevant (e.g. dispensary fee).
- Some medical orders given at high frequency during treatment.
  If several medical orders appear in a sequence with high frequency and has a high average frequency in the sequence, there is a possibility that some important medical orders may not appear. This may be because many medical orders that appear frequently and have a high average frequency in the sequence may obscure other medical orders. Therefore, it is necessary to exclude medical orders that appear frequently and have a high average frequency in the sequence (e.g. percutaneous oxygen saturation measurement). $TO/S(term\ occurrence\ per\ sequence)$ is an indicator that can represent the average number of occurrences of certain medical orders, which can be used to determine the medical orders to be excluded.

$$TO/S = \frac{Total\ number\ of\ order\ occurrences}{Total\ number\ of\ sequences}$$

B. Extraction of frequent medical-order patterns for each infection wave

To show the differences in frequently occurring medical-order patterns by infection waves, we first apply SPM to a set of EMR data for the target disease at each medical institution for different periods and derive the SVs for each infection wave. Next, LCSVs, which are the common parts among the SVs, are extracted. The LCSVs and the original SVs are then used to derive the labeled (colored) MSV for each infection wave to show the differences among the periods.

C. Extraction of SVs with temporal characteristics by applying a sliding window

To extract and visualize frequent patterns with different characteristics at different times, we adopt the concept of the sliding window, which requires choosing an appropriate window type, the window size, and the slide unit for the window.

Three types of sliding window are proposed: the time interval window, the sequence number window, and the time-sequence number window.

1) Time interval window: The window size is set to a specific time range, and within that range, SPM is applied to extract frequent patterns, such as over a period of one week or one month. The window is then moved by the set slide unit and SPM is reapplied to the data in the new window.

However, the time interval window has one drawback: for periods with a small number of sequences, applying SPM may not extract anything. Therefore, in this paper, we propose using the sequence number window.

2) Sequence number window: Here, the sequence data in the database are sorted in time order, and then SPM is applied to sequences of the size of the window to extract the frequent patterns.

The sequences before and after the peak period in the infection wave are considered to have the most temporal characteristics. Using these sequences, the characteristic patterns during the peak period in the infection wave can be identified.

To extract frequent patterns with different timing characteristics, the standard deviation is calculated from the data before and after the peak period in the infection wave, and a doubled standard deviation containing approximately 95% of the data on one side is adopted. The fourfold standard deviation, which is the width of the peak, is then used as the window size for that period, which is sufficient to enable the extraction of frequent patterns that have the characteristics of that period. Here, given that the window size must be an integer, the value of the window size calculated from the standard deviation is rounded to the nearest integer.

Because the number of hospitalized patients converges at the beginning and end of each infection wave, the sequence data for the corresponding period also decrease. If the slide unit is large, the sequence at the beginning of the next infection wave will be skipped, which may prevent us from extracting medical-order patterns with different temporal characteristics at the beginning of the wave. Therefore, this paper proposes using the time-sequence number window to dynamically change the slide unit.

3) Time-sequence number window: The time-sequence number window is based on the sequence number window, with the number of sequences in subsequent periods being considered.

The approach to changing the slide unit dynamically can be described as follows.

1) First, the default slide unit must be set. There are various ways to initialize the slide unit, but in this paper, the slide unit was set to half of the window size for the evaluation experiments. Here, if the calculated slide unit is not an integer, it is rounded to the nearest integer.
2) Then the number of sequences in the set time parameter range from the beginning of the window must be determined.
3) If the number of sequences is smaller than the default slide unit, the slide unit is made equal to the number of sequences. Otherwise, the slide unit is set to a default value.
4) Finally, the window by the slide unit is moved, and SPM is reapplied to the data in the new window.

D. Comparison of Frequent Patterns Between Windows

By comparing adjacent windows one by one, it is possible to detect when there has been a major switch in treatment patterns.

LCSV is an extension of the idea of the longest common subsequence (LCS), enabling a similar extension of the comparison method between LCSs. This gives comparisons between LCSVs using the following $SVS(Sequence\ Variant\ Similarity)$ formula.

$$SVS(SV1, SV2) = \frac{2|LCSV|}{|SV1| + |SV2|}, (0 \leq SVS \leq 1)$$

Here, $|LCSV|$, $|SV1|$, and $|SV2|$ are the number of nodes in LCSV, SV1, and SV2, respectively. The SVS value is an indicator of the similarity of the treatment patterns between the two groups. When the SVS value is high, the treatment patterns of the two groups are similar, and, conversely, when the SVS value is low, the treatment patterns of the two groups are substantially different.

IV. Experiment

We performed an experimental evaluation of our proposed method to visualize the differences in medical orders between different infection waves and the transition state of medical orders using a real COVID-19 dataset.

A. Target data

In our experiments, we used medical-order data from the actual EMR data for COVID-19 provided by two sources (Medical Institutions A and B), which were collaborating organizations in the "Survey of the Impact of COVID-19 Infection on Medical Practice and Development of a Predictive Model" [8].

The data were divided according to the period of hospitalization corresponding to each wave of the COVID-19. The number of patients and sequences in each wave from Wave 1 to Wave 5 for Medical Institutions A and B is shown in Table I. The number of sequences is greater than the number of patients because the same patient may be involved in more than one sequence.

TABLE I
NUMBER OF PATIENTS AND SEQUENCES PER WAVE OF COVID-19

	Medical Institution A		Medical Institution B	
	Patients	Sequences	Patients	Sequences
First wave (2020.04.01–2021.06.30)	42	49	10	10
Second wave (2020.07.01–2020.10.31)	52	63	35	35
Third wave (2020.11.01–2021.02.28)	96	108	55	57
Fourth wave (2021.03.01–2021.06.30)	89	99	24	24
Fifth wave (2021.07.01–2021.09.30)	98	104	35	35

B. Setup

For the proposed method, as mentioned in Section II, it is necessary to exclude medical orders that had little relevance to treatment and those that were always given frequently to extract significant SVs. After excluding medical orders that had little relevance to treatment, the average number of occurrences of each medical order was obtained by calculating the TO/S values of all medical orders present in the SDB. Next, medical orders that were always given frequently were

excluded on the recommendation of the medical staff. As a result, our experiments excluded medical orders with TO/S values of 1.17 or higher for Medical Institution A and 1.04 or higher for Medical Institution B.

To be able to apply SPM and extract characteristic medical orders, an appropriate minimum support value (minsup) should be set [1]. We experimented with several minsup values and extracted SVs to check the contents. We settled on values of 0.2 for Medical Institution A and 0.25 for Medical Institution B for the extraction of frequent patterns for each infection wave. For the minsup in the sliding window applied to derive the transition history of medical orders, it was decided to adopt 0.2 for both institutions. If the minsup adopted were too high or too low, medical orders with relevant characteristics would not be extracted.

The names of the medical orders and the abbreviations used in the figures presenting experimental results are shown in Table II.

TABLE II
ABBREVIATIONS FOR NAMES OF MEDICAL ORDERS USED IN THIS PAPER

Abbreviation	Medical order name	Abbreviation	Medical order name
AT	Acetaminophen	MAT	Magmitt Tablets
AL	Alvesco	MOT	Magnesium Oxide Tablets
AS	Aspirin	MT	Medicon Tablets
CM	Camostat mesilate	MTT	Metoclopramide Tablets
CTD	CT diagnostics	OT	Oxygen Therapy
DAT	Daiphen Tablets	PCT	Patent Cooperation Treaty
DT	Decadron Tablets	PT	Prednisolone Tablets
FRTN	Ferritin	RS	Risperidone
HBsAg	Hepatitis B surface antigen	ST	Sennoside Tablets
LC	Lyrica Capsules	TKD	Tsumura-Kampo Daikenchuto
LS	Loxoprofen Sodium Hydrate	VWFAg	Von Willebrand factor antigen
LT	Lunesta Tablets	XT	Xarelto Tablets

C. Extraction of SVs by infection wave

Using the EMR data from Medical Institutions A and B, the SVs were obtained by applying SPM to the medical-order sequences for each infection wave, after excluding the medical orders with high TO/S values and orders that had little relevance from the medical-order data of patients .

D. Derivation of MSV for different infection waves

For the MSV shown in Figure 1, the LCSV between the fourth and fifth waves in Japan is shown in yellow, the fourth wave in blue, and the fifth wave in red. By labelling the LCSVs and their respective infection waves with corresponding colors, it is possible to visualize the differences in the frequent medical-order patterns for the adjacent infection waves.

Figure 1 shows that DT appeared 9 times(Day 3 - Day 11) in the fourth wave of SV, while in the fifth wave, DT appeared only 4 times(Day 5 - Day 8). Therefore, DT use was reduced in the fifth wave in Japan compared with the fourth wave. Based on feedback from healthcare professionals, it is believed that the reduction in the number of cases of severe pulmonary injury caused by COVID-19 may be due to the vaccine's effectiveness among the elderly and the administration of neutralizing antibody drugs [6]. The fifth wave also shows a branch between with and without OT, while in the fourth wave there is no such branch. In the branch requiring OT, steroid DT use and follow-ups were performed, whereas for the branch not requiring oxygen inhalation, the MT cough suppressant was administered, which was a frequent pattern.

E. Differences between the medical institutions

Even during the same infection wave, the pattern of frequent medical orders differed between medical orders. The MSVs for Medical Institutions A and B during the third wave in Japan are shown in Figure 2. For the MSVs in Figure 2, since the number of nodes in lcsv is only 2 and most of the other medical orders are different,it is clear that the frequent medical-order patterns for the two sets of medical orders in the same third wave in Japan were almost different. In particular, for the administration of medical orders, Medical Institution A used CM, whereas Medical Institution B used XT. For Medical Institution B, steroids such as DT did not appear in the pattern of frequent medical orders, and the medical order for the administration of MTT, a drug for the treatment of gastrointestinal dysfunctions, appeared as a frequent medical-order pattern in the early stages of admission.

Based on feedback from healthcare professionals, it is believed that the differences in treatment policies between medical institutions have resulted in different results.

F. Analysis of the transition history of medical orders

In our experiments, three types of sliding window were implemented to extract patterns with temporal characteristics.

Because the sequence number window cannot completely extract patterns with temporal characteristics, experimental results with the sequence number window are not presented in this paper.

The time interval window was applied to the data from the first to the fifth wave in Japan for Medical Institution A, and the resulting transition diagram is shown in Figure 3. The window size was set to one month and the slide unit was set to one week.

The time-sequence number window was applied to the data from the first to the fifth wave in Japan for Medical Institution A, and the resulting transition diagram is shown in Figure 4. To extract frequent patterns with different temporal characteristics, four times the standard deviation was used as the window size for that period, and the default slide unit was set to half of the window size for the evaluation experiment.

In the transition diagram using the time interval window shown in Figure 3, the first wave is shown in purple, the second wave in light blue, the third wave in green, the fourth wave in blue, and the fifth wave in red. Using information from medical guidance about COVID-19 [7], the peak periods of the number of infections in all five waves and the start dates for the first and second vaccinations are indicated by

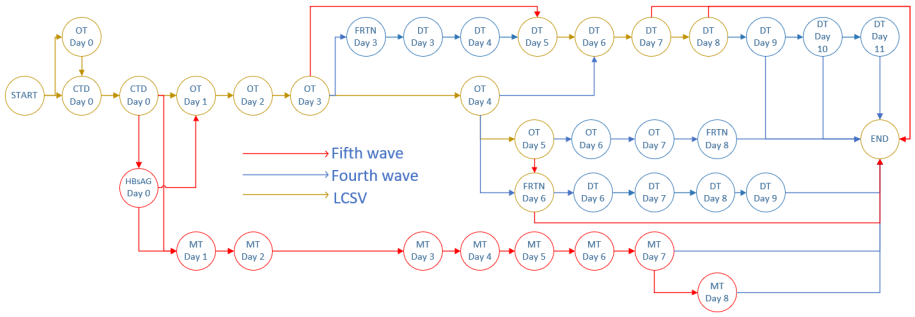

Fig. 1. MSV for the fourth and fifth waves in Medical Institution A

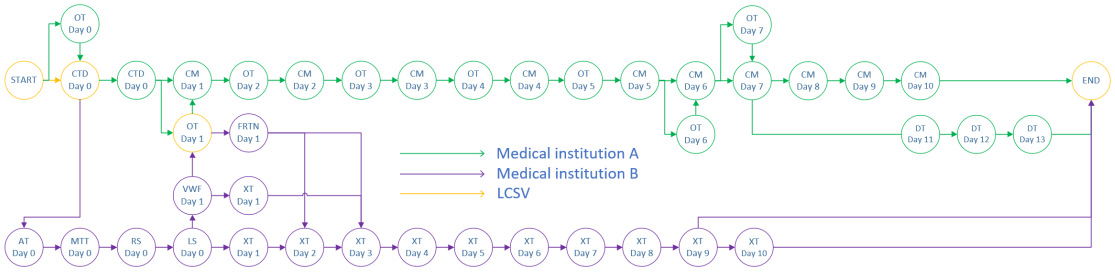

Fig. 2. MSV for the third wave in Medical Institutions A (minsup=0.2) and B (minsup=0.25)

brackets and arrows in Figure 3. In addition, the figure shows the increase or decrease in medication and OT use and the change in inspection-only pattern and medication when the pattern changes significantly.

A closer look at the transition diagram for the time interval digital window in Figure 3 shows that CM was available during the first-wave peak, but, subsequently, it was no longer used and replaced with TKD, MAT, and LC. However, in the second-wave peak, CM was again used as the main drug. After the peak of the third wave, CM was not used after the start of the first and second vaccinations. Conversely, the steroid DT appeared in the fourth wave, but the peak of the fourth wave showed the appearance of SM. After that, the peak of the fifth wave showed the appearance of MT. Note also that the pattern can change within the same wave of infection.

Based on feedback from healthcare professionals, it is believed that during the first to third waves, CM was used because it was reported to be effective in treating COVID-19. However, CM was stopped in June 2021 because it was judged to be ineffective.

G. Differences between the time interval window and the time-sequence number window

Even for the same infection wave, the transition diagram appears to differ, depending on the type of window used. For example, using the time interval window, the medications TKD, MAT, LC, and OB appeared in the first and second waves. This was not the case when using the time-sequence number window. Using the time-sequence number window, the major medication transitions are the same as when using the time interval window, but the frequency of change is less for increases and decreases in medication and the usage of OT. In addition, the medication types that appear are fewer than when using the time interval window.

We can summarize the differences as follows. Using the time interval window, medication patterns changed more frequently, and a wider variety of medications emerged. By contrast, using the time-sequence number window, changes are less frequent, so the changes in the use of major medications at each time period are more readily apparent.

Therefore, the time interval window should be used to capture changes in medication usage at each time period and to capture more detailed medication use, whereas the time-sequence number window should be used to capture the major medication transitions at each time period.

V. Conclusion

In this paper, LCSV- and MSV-based techniques were used to analyze the differences between frequent medical-order patterns of the same disease during different periods. The SVs for each period were extracted from the EMR data from the first to the fifth wave of COVID-19 in Japan, and

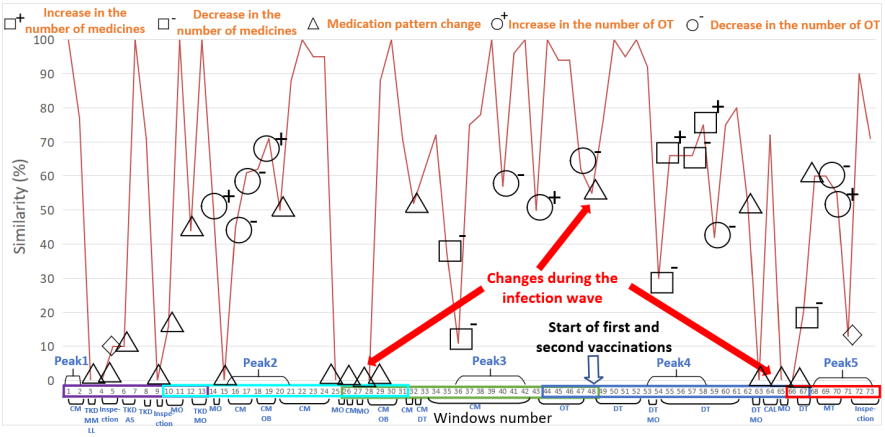

Fig. 3. Similarity transition diagram between time interval windows for Medical Institution A

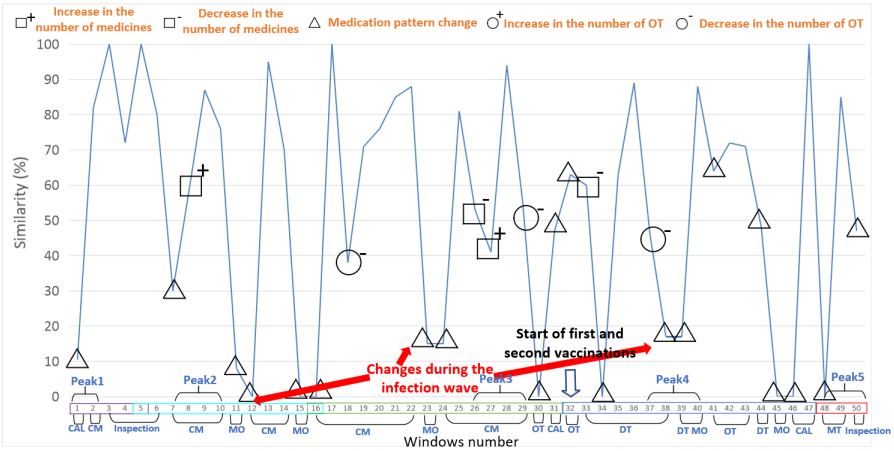

Fig. 4. Similarity transition diagram between time-sequence number windows for Medical Institution A

MSVs were obtained by fusing the SVs for each infection wave with LCSVs to visualize the differences in treatment patterns between infection waves. Frequent patterns with different temporal characteristics were extracted by applying and adjusting three types of sliding window, in terms of window sizes and slide units. The frequent SVs of adjacent windows were evaluated using LCSVs to calculate similarities for use in analyzing the transition history of medical orders. Positive comments from medical staff have confirmed that the visualization would benefit the comparison of treatment patterns between time periods and be useful in the application of medical assistance.

ACKNOWLEDGEMENT

The research is partially supported by Grants-in-Aid from Japan Science and Technology Agency (20H04192, 21K1774), and uses EMR data provided by collaborating organizations in the project of Survey of the Impact of COVID-19 Infection on Medical Practice and Development of a Predictive Model supported by Kansai Economic Federation.

REFERENCES

[1] R. Agrawal, R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases. Proceeding of the 20th International Conference on Very Large Databases,” 1994.

[2] H. H. Le, H. Edman, Y. Honda, M. Kushima, T. Yamazaki, Kenji Araki, H. Yokota, “Fast Generation of Clinical Pathways Including Time Intervals in Sequential Pattern Mining on Electronic Medical Record Systems”. Proceeding of the fourth International Conference on Computational Science and Computational Intelligence (CSCI 2017), pp. 1726-1731, 2017.

[3] Y. Honda, M. Kushima, T. Yamazaki, K. Araki, H. Yokota, “Detection and Visualization of Variants in Typical Medical Treatment Sequences. Proceeding of the 3rd VLDB Workshop on Data Management and Analytics for Medicine and Healthcare,” Springer, pp. 88-101, 2017.

[4] Y. Li, H. H. Le, R. Matsuo, T. Yamazaki, K. Araki, H. Yokota, “Comparison of Sequence Variants and the Application in Electronic Medical Records,” Proceeding of the 33rd International Conference on Database and Expert Systems Applications (DEXA2022), Vol. 13427, pp. 117-130, 2022.

[5] H. H. Le, Y. Horino, K. Araki, T. Yamazaki, H. Yokota, “Sequential Pattern Mining of Large Combinable Items with Values for a Set-of-items Recommendation,” , Proceeding of the 34th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2021), pp. 56-61, 2021.

[6] Characteristics of the 5th wave (in Japanese) https://web.pref.hyogo.lg.jp/governor/documents/g_20210728_01_02.pdf

[7] Guideline for the Treatment of New Coronavirus Infections COVID-19, Edition 7.0 (in Japanese) https://www.mhlw.go.jp/content/000904136.pdf

[8] Survey of the Impact of COVID-19 Infection on Medical Practice and Development of a Predictive Model (in Japanese) https://www.nagoya2.jrc.or.jp/content/uploads/2021/08/8125.pdf