Statistical Analysis and Visualization of Medical Instruction Patterns in Multimedia-Rich Electronic Medical Records

Miwa Sugitani[1], Ryosuke Matsuo[1], Tomoyoshi Yamazaki[1], Kenji Araki[1],
Masato Oguchi[1], Haruo Yokota[2], and Hieu Hanh Le[1]
[1]Ochanomizu University, Tokyo, Japan, [2]Josai University, Tokyo, Japan
E-mail: {g2120520, oguchi, le}@is.ocha.ac.jp, matsuo@ldi.or.jp, {yamazaki.cp, araki6925, yokota.h.aa}@gmail.com

***Abstract*—While the adoption of multimedia-rich electronic medical records (EMRs) has advanced the analysis of clinical data, differences in data formats and notations across institutions render inter-institutional comparison and feature extraction difficult under conventional methods based on single-institution data. Moreover, although extracting disease-specific instruction patterns from multi-institutional data is valuable, the analysis becomes complex due to variations in treatment policies. This paper proposes a method for integrating data from multiple institutions using data standardization and entity resolution techniques, enabling analysis based on a unified framework. Disease-specific instruction sequences are extracted from EMRs, and frequent closed patterns are identified by considering time intervals based on date information. By retaining sequence identifiers throughout the analysis, it becomes possible to reference the corresponding patients' test results. Evaluation experiments using real EMR data from 27 institutions were conducted. The results demonstrate the effectiveness of the proposed method by visualizing branching structures in frequent instruction sequences, related statistical information, and the fluctuation of abnormal test result rates, thereby confirming its utility for decision support and improving clinical processes.**

***Index Terms*—electronic medical record data, sequential pattern mining, visualization**

I. Introduction

In recent years, the adoption of electronic medical records (EMRs) has expanded from large hospitals to smaller institutions, accelerating the digitalization of health care. Initiatives such as the "Millennium Medical Record Project" [1] have also promoted the aggregation and secondary use of medical data, opening up new possibilities for clinical support and research.

EMR data have become a fundamental resource to help derive health care insights and provide more effective health care through applying data engineering techniques for analysis. These multimedia-rich data comprise records of hospital visits from patients, typically involving heterogeneous medical instructions, such as diagnoses, tests, prescriptions, injections, and vital signs together with the activation time of these orders. As such, data-mining techniques, including sequential pattern mining (SPM), can be applied effectively to the analysis of such EMR data.

The analysis of medical instruction sequences contributes to the standardization and improvement of clinical workflows, enhancing both the consistency of medical quality and patient outcomes. At the same time, personalized support tailored to individual patient conditions has been increasingly emphasized [2]. Furthermore, utilizing data from other institutions can help improve internal clinical practices, highlighting the growing importance of inter-institutional data sharing and analysis.

However, many existing studies are based on data from a single medical institution, and thus inter-institutional studies of clinical processes are lacking [3]–[6]. On the other hand, there have been another previous studies that focused on single-disease datasets across institutions [7], [8]. In such studies, the extraction of statistical information is inadequate, and its relationship with medical instruction patterns is not visualized effectively, which limits its use in clinical decision-making. In sum, while existing works depict the branching of frequent medical instruction patterns, they do not incorporate the underlying statistical information or abnormal test values.

In this study, we propose a method for integrating EMR data from multiple medical institutions, extracting disease-specific frequent instruction patterns with useful insights, such as statistical information and the distribution of normal and abnormal test results. The proposed method adopts the T-PrefixSpan algorithm [9] to capture the temporal characteristics of clinical processes by considering the time intervals between the instructions. Through this approach, our goal is to visualize the inter-institutional differences in clinical workflows and contribute to the standardization and quality improvement of health care.

Experimental evaluation using anonymized real-world EMR data from 27 institutions focus on three aspects: extraction of patterns with associated statistical information, analysis of normal and abnormal test result rates, and visualization of differences in frequent instruction patterns. The experimental results demonstrate the usefulness of our method in supporting clinical decision-making and improving health care processes.

The remainder of this paper is organized as follows. Section II describes the background knowledge, and related technologies and studies. Section III presents the proposed method. Section IV reports the experimental results. Finally, Section V

concludes the paper and discusses future work.

II. Background Knowledge and Related Works

This section describes the background knowledge about sequential databases and related works on sequential pattern mining.

A. Background Knowledge

Here, we show basic definitions used in sequential pattern mining. A sequence is defined as an ordered set of items, and the sequential database (SDB) consists of elements that pair each sequence with a unique sequence identifier (SID).

Definition 1 (Sequence): Given an item set $I = \{i_1, i_2, \ldots, i_n\}$, a sequence S is defined as:

$$S = (\{s_1, s_2, \ldots, s_m\}, \prec_S)$$

where $s_j = (id, i),\ i \in I$ represents the item and id is its index in the sequence. The relation $\prec_S$ denotes a total order over S, satisfying:

$$\forall s_i, s_j \in S,\ s_i \prec_S s_j \vee s_j \prec_S s_i$$

Definition 2 (Sequential Database): An SDB is defined as:

$$SDB = \{(sid_1, S_1), (sid_2, S_2), \ldots, (sid_m, S_m)\}$$

where sid_i uniquely identifies sequence S_i.

Definition 3 (Frequent Sequence): A sequence in the SDB that satisfies a minimum support threshold minSup is called a frequent sequence.

Definition 4 (Time Sequence): Given an item $i \in I$ that occurs at time t, the time item is expressed as (i, t). A time sequence is an ordered list of such time items:

$$s = \langle (i_1, t_1), (i_2, t_2), \ldots, (i_n, t_n) \rangle$$

where each (i_k, t_k) represents the item and its timestamp.

Definition 5 (Sequence Variant (SV)): Given a time sequence s, a sequence s' with the same item set I but different time intervals or order is called a sequence variant. The set of sequence variants is defined as:

$$SV(s) = \{s' \in SDB \mid s' \neq s\}$$

Here, s' is a sequence in the SDB that is related to s, and SV captures the structural and temporal diversity of frequent sequences.

Definition 6 (Closed Pattern): A pattern is a closed pattern if there is no supersequence with the same support as it and containing it

B. Related Works

This section introduces studies related to the present work. That have contributed to the standardization and quality improvement of clinical processes and serve as a foundation for our research.

1) Sequential Pattern Mining: SPM is a very active research topic that includes numerous extension approaches for specific needs. This section introduces major approaches related to this work: frequency-based SPM and time-interval SPM. A well-known SPM algorithm is the Apriori-based frequent-pattern mining algorithm [10]. However, it is very time-consuming with large datasets and generates many irrelevant patterns among its results. To exclude patterns, PrefixSpan [11] was proposed to mine the complete set of patterns while reducing the effort of candidate pattern generation by exploring prefix projection. To improve efficiency further, CSpan [12] was proposed for mining closed sequential patterns. For a sequence seq, if there is no supersequence with the same support as it and containing it, seq is a closed sequence.

In addition to constraining the format of sequential patterns to reduce the number of sequential patterns obtained by mining, some algorithms extract the sequential patterns that match the users' requirements, such as Top-k SPM [13]–[16] and targeted SPM [17]–[19].

Initially, the method proposed by Agrawal et al. [10] did not consider the time interval between items. For example, a sequence consisting of a prescription followed by surgery the next day was considered the same as a sequence that performed surgery two days after the prescription. Chen et al. proposed a mining method called TI-SPM for sequences where the time interval is important, such as medical instructions, which should treat the presented two sequences as different [20].

T-PrefixSpan [9] is a method of extracting frequent sequential patterns from EMR data that considers time intervals and the efficacy of medicines. T-CSpan [21] further improves the speed performance by applying the idea of mining only closed patterns.

2) Frequent Instruction Pattern Analysis from the Data of a Single Institution: Honda et al. detected the common part of the closed frequent pattern with the same number of items for each relative treatment day [3]. Then, Le et al. proposed methods to evaluate the reasons why variants might appear by using multivariate analysis [6], [7]. Some studies have focused on performing SPM for a set of item recommendations from tens of thousands of items with values in a very large sequential database [4]. The SPM method first vectorizes a combination of items and item values and identifies clusters of item-set types. From the item-set types that best suit the target input, a set of items is recommended, in terms of both frequency and uniqueness.

However, that study was limited to a single institution, and therefore did not support inter-institutional comparisons or standardization. In contrast, our study aims to integrate data from multiple institutions and extract common frequent patterns to enable such comparisons and promote standardization and collaboration.

3) Frequent Instruction Pattern Analysis Across the Data of Multiple Institutions: One method has compared two SVs by proposing the concepts of common SVs, the longest common sequence variants, and the merged sequence variants and

describing the algorithms to calculate these sequences [22]. Based on that idea, Zhao et al. proposed a method to analyze the transitions in differences between medical orders for the treatment of COVID-19 [5].

Some work has also analyzed EMR data related to COVID-19 across multiple institutions, extracting frequent instruction patterns, and clustering them based on sequence similarity [8]. The similar patterns within clusters are merged to identify common and different instruction flows, contributing to standardization and inter-institutional collaboration.

However, that study focused on a single disease, and did not address analyses involving multiple diseases or comparisons of cross-disease instruction flows.

III. Our Proposal

This study aims to support the standardization of clinical processes and assist medical professionals in decision-making by extracting frequent medical instruction patterns with statistical information from EMR data and associating them with normal and abnormal test result occurrences. The proposed method consists of the following four steps: Construction of medical instruction sequences, Extraction of frequent medical instruction patterns with statistical information, Calculation of normal and abnormal test results rates and Visualization.

A. Construction of Medical Instruction Sequences

Organizing patient treatment records in chronological order is essential for analyzing clinical workflows and promoting standardization. In this study, we collect each patient's treatment history from EMR data and construct a medical instruction sequence by arranging medical instructions in temporal order.

The day on which a medically significant instruction (e.g., surgery) was performed is set as the reference date (Day 0), and other instructions are arranged according to their relative elapsed days. If the same patient is admitted multiple times, each admission is treated as an independent clinical process.

Definition 7 (Ordering Rule for Same-Day Instructions): Assuming no timestamp is available, if multiple instructions occur on the same day, they are ordered as follows:

Surgery → Medication → Test → Medical procedure

Within the same category, instructions are further sorted based on a predefined dictionary in lexicographic order. This rule helps standardize the sequence order, improving the accuracy of pattern extraction.

Definition 8 (Medical Instruction Sequence): The medical instruction sequence for a patient p_i is defined as:

$$S_{p_i} = \langle s_1, s_2, \ldots, s_n \rangle$$

Each s_j represents the j-th instruction and is expressed as $s_j = (t_j, a_j)$, where t_j is the elapsed day and a_j is the instruction type (e.g., surgery, medication, test).

Definition 9 (Elapsed-Day Medical Instruction Sequence): Let t_0 be the reference date (e.g., surgery day). Each instruction is then represented with the elapsed day $d_j = t_j - t_0$:

$$S_{p_i} = \langle (d_1, a_1), (d_2, a_2), \ldots, (d_n, a_n) \rangle$$

If the patient has no surgery, the admission date is used as the reference. Multiple instructions may share the same d_j.

The resulting instruction sequences are stored in a new table with a unique sequence identifier (SID) per patient. These sequences allow for trend analysis and provide informative reference points for clinical decision-making.

B. Extraction of Frequent Medical Instruction Patterns With Statistical Information

From the constructed sequences, we extract frequent disease-specific instruction patterns. We adopt the T-PrefixSpan algorithm [9], which considers both the order and time intervals between instructions.

T-PrefixSpan extends PrefixSpan [23] by addressing limitations of I-PrefixSpan [9], enabling more flexible interval handling. While conventional PrefixSpan only considers order, T-PrefixSpan captures time dependencies between instructions.

Definition 10 (Medical Sequential Database (MSDB)): A medical sequential database (MSDB) is composed of instruction sequences for multiple patients:

$$D = \{(SID_1, S_{p_1}), (SID_2, S_{p_2}), \ldots, (SID_m, S_{p_m})\}$$

Here, SID_i is the sequence ID for patient p_i and S_{p_i} is their instruction sequence.

Definition 11 (Frequent Sequence with T-PrefixSpan): A time-interval-aware frequent sequence fs is expressed as:

$$fs = \langle (a_1, x_1), (a_2, x_2), \ldots, (a_n, x_n) \rangle$$

where $x_j = t_{j+1} - t_j$ denotes the interval between instructions a_j and a_{j+1}. T-PrefixSpan enables the extraction of time-sensitive frequent patterns that capture the rhythm and structure of clinical processes.

Using this approach, we accurately identify significant patterns in real clinical workflows, including time dependencies. For each extracted pattern, we then collect test results performed during those instructions and calculate summary statistics, such as mean, median, min, and max.

Definition 12 (Frequent Pattern with Statistical Information): For a frequent instruction pattern fs, the associated test result set $X(\alpha)$ for test α is:

$$X(\alpha) = \{x_1, x_2, \ldots, x_n\}$$

The statistical information is computed as follows:

$$\begin{aligned}
&\text{Mean:} && \mathrm{Mean}(X(\alpha)) = \tfrac{1}{k} \textstyle\sum_{j=1}^{k} x_j \\
&\text{Median:} && \mathrm{Median}(X(\alpha)) = x_{\lceil \frac{k}{2} \rceil} \\
&\text{Max:} && \mathrm{Max}(X(\alpha)) = \max(X(\alpha)) \\
&\text{Min:} && \mathrm{Min}(X(\alpha)) = \min(X(\alpha))
\end{aligned}$$

C. Calculation of Normal and Abnormal Test Result Rates

For each frequent instruction pattern of a specific disease, we calculate abnormal test result rates per elapsed day. Sequences are split into two groups: those containing the pattern (D_{in}) and those not containing it (D_{out}):

$$D_{\text{in}} = \{S_{p_i} \in D \mid fs \subseteq S_{p_i}\}, \quad D_{\text{out}} = D \setminus D_{\text{in}}$$

Test results for each elapsed day d_j are collected and classified as:

$$C(x_j) = \begin{cases} -1 & (x_j < B(\alpha)) \quad \text{(Low)} \\ 0 & (x_j \in [B(\alpha), T(\alpha)]) \quad \text{(Normal)} \\ 1 & (x_j > T(\alpha)) \quad \text{(High)} \end{cases}$$

The rates are calculated as follows:

$$HR(\alpha, d_j) = \frac{|\{x_i \in X(\alpha, d_j) \mid C(x_i) = 1\}|}{|X(\alpha, d_j)|} \quad \text{(High rate)}$$

$$LR(\alpha, d_j) = \frac{|\{x_i \in X(\alpha, d_j) \mid C(x_i) = -1\}|}{|X(\alpha, d_j)|} \quad \text{(Low rate)}$$

$$NR(\alpha, d_j) = 1 - (HR(\alpha, d_j) + LR(\alpha, d_j)) \quad \text{(Normal rate)}$$

Here, $X(\alpha, d_j)$ represents the set of test results for test α at elapsed day d_j.

D. Visualization

We propose a visualization method that represents frequent medical instruction patterns and normal and abnormal test result rates to reveal structural features and branching points in clinical workflows. Each node represents a medical instruction, and nodes within the same pattern are connected by edges. Nodes are uniformly sized and arranged vertically based on their support values, with more frequent nodes positioned higher to emphasize important patterns. Abnormal test result rates are visualized using bar charts. When hovering over a bar, a tooltip displays detailed information, and the corresponding nodes are highlighted, enabling intuitive identification of patterns with concentrated abnormalities. Additionally, by filtering specific test items, users can focus on abnormalities and analyze temporal factors, such as the days when abnormalities are more likely to occur, supporting further analysis of causative factors.

This visualization tool effectively supports a deep understanding of clinical processes and risk points, serving as a practical decision support system for health care professionals.

IV. Experimental Evaluation

This section describes the objectives of the experiment, the details of using the dataset, the experimental environment, the experiment method, and finally the experimental results.

A. Objectives

The objective of this evaluation is to verify the effectiveness and feasibility of the proposed method from the following perspectives.

1) **Visualization of Variants in Frequent Instruction Patterns.** The extracted frequent instruction patterns are validated by an observation of whether they involve basic instructions for the targeted disease treatment. Moreover, we evaluate whether differences between frequent medical instruction patterns (variants) can be appropriately visualized. The goal is to enable medical professionals to intuitively grasp the branching points and anomalies in clinical flows.

TABLE I
Patient count per disease in the real dataset

Disease	No. of Patients
Malignant lung tumors	9,962
Acute myocardial infarction	4,874
Angina pectoris and chronic ischemic heart disease	11,148
Malignant neoplasm of stomach	5,927
Malignant neoplasm of liver and intrahepatic bile ducts	2,030
Degeneration and hernia of intervertebral discs	1,194
Malignant neoplasm of breast	12,818
Bladder tumor	9,748
Malignant neoplasm of cervix and uterus	897
Total	58,598

TABLE II
Experimental Environment

Programming language	Python
Database	PostgreSQL ver. 16.6
Database driver	Psycopg ver. 3

2) **Extraction of Frequent Instruction Patterns with Statistical Information.** We assess whether frequent medical instruction patterns with accompanying statistical information can be accurately extracted through sequence mining.
3) **Calculation of Normal and Abnormal Value Occurrence Rates in Test Results.** We verify whether the abnormal value occurrence rates in test results associated with frequent instruction patterns can be correctly calculated. Furthermore, we compare the abnormality rates between patients who follow the frequent patterns and those who do not, to identify characteristic tendencies in each group.

B. Dataset

The dataset used in this experiment contains 58,598 records of nine diseases from 27 medical institutions, including malignant lung tumors and ischemic heart diseases, etc. The present experiment was approved by the Ethics Committee of the Life Data Initiative (No 2024_MIL_0004_A001).

Table I shows the dataset's number of patients per disease.

C. Experimental Environment

In this experiment, the PrefixSpan library[1] was modified to implement functions equivalent to T-PrefixSpan to extract frequent patterns. The dataset was stored in a PostgreSQL database and accessed via Python using the Psycopg library, which enabled efficient processing of large-scale queries. This setup allowed effective handling of multiple diseases and medical instruction sequences to improve the performance of frequent pattern extraction.

1) Frequent Instruction Pattern Extraction: We modified the PrefixSpan library [23] to implement T-PrefixSpan [9] functionality. T-PrefixSpan treats the combination of elapsed days from the reference date and medical instructions as a

[1]https://github.com/chuanconggao/PrefixSpan

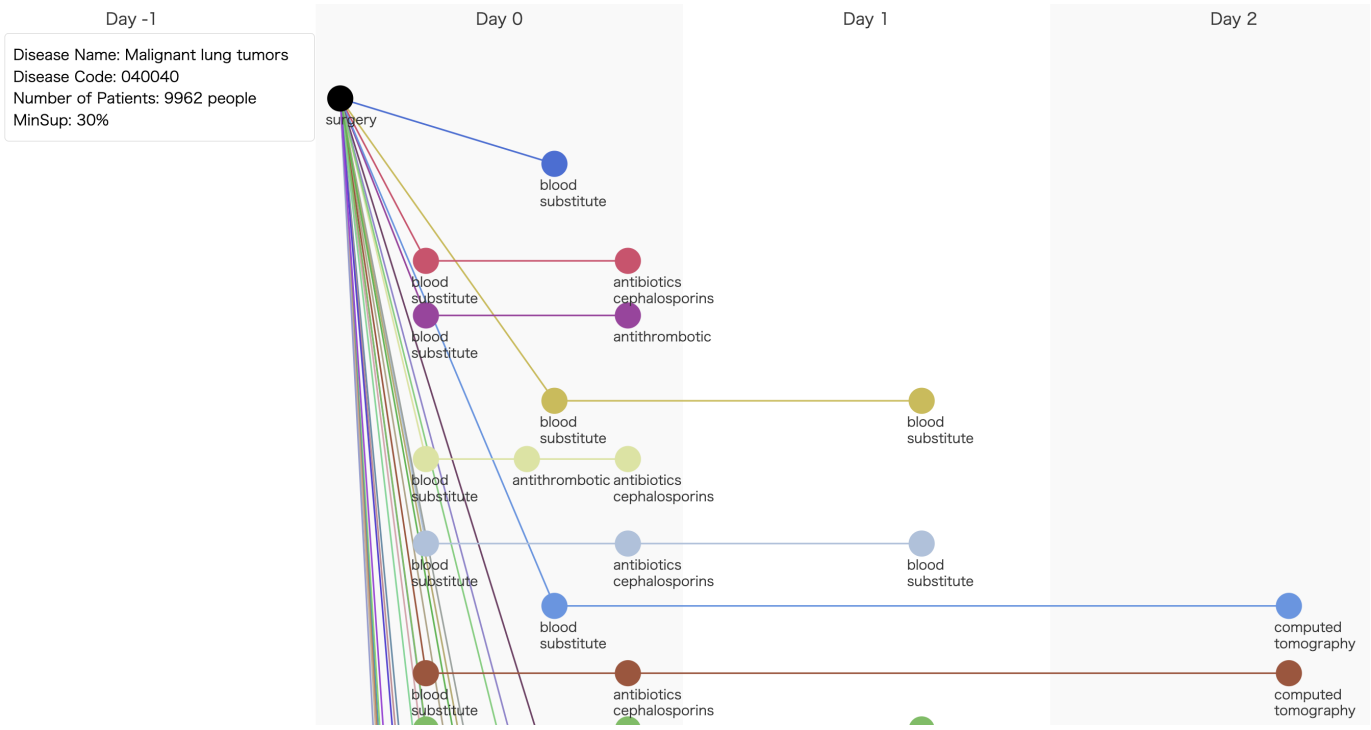

Fig. 1. Visualization of frequent medical order patterns

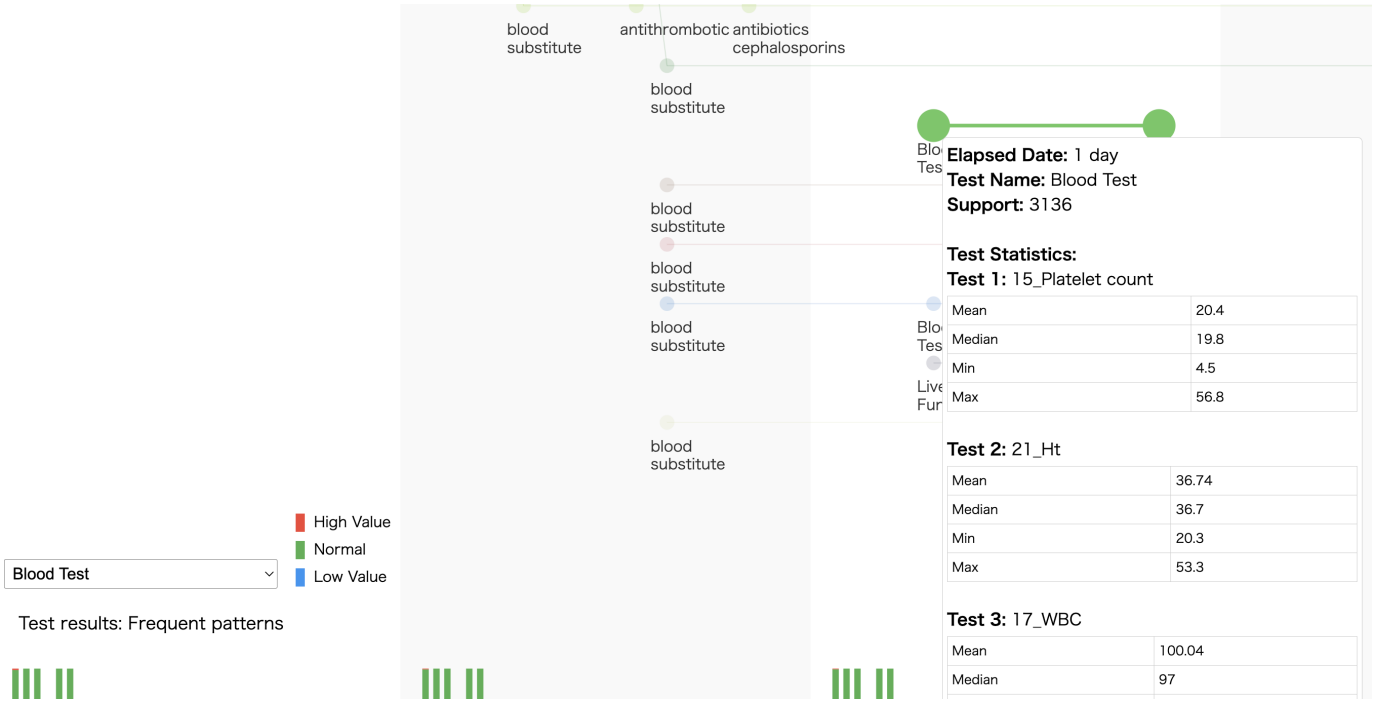

Fig. 2. Node-clicking statistics and highlighting

single item, enabling pattern extraction that retains temporal information. For each ICD-10 code group, frequent medical instruction patterns were extracted. A minimum support threshold (minSup), defined as a proportion of total sequences per disease, was applied to filter patterns.

D. Experimental Method

The experiment follows these steps: (1) generate disease-specific instruction sequences from the dataset, (2) extract frequent patterns using T-PrefixSpan, (3) calculate test result statistics and abnormality rates by elapsed day, (4) visualize

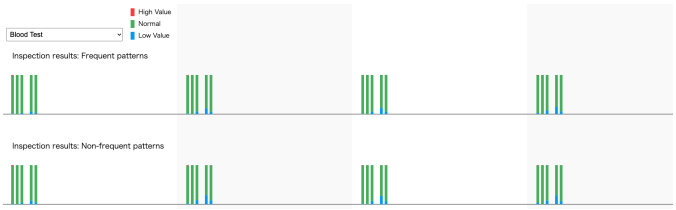

Fig. 3. Visualization of the incidence of abnormal values in instruction results (default state)

Fig. 4. Visualization of the rate of abnormal values in instruction results (hover state)

these results and assess whether differences and critical points in clinical processes are clearly represented. This allows a comprehensive evaluation of the proposed method's contribution to clinical process analysis and improvement.

E. Experimental Results

This section presents the visualized results of the extracted frequent instruction patterns and abnormal value occurrence rates.

1) Visualization of Frequent Instruction Patterns: As an example, Fig. 1 visualizes frequent medical instruction patterns for 9,962 patients with malignant lung tumors. The patterns were extracted with a minimum support threshold (MinSup) of 30%. Each node represents a medical instruction, and the edges indicate their sequential relationships. Patterns with higher support values are positioned toward the top. The visualization confirms that the most frequent patterns involve both the administration and non-administration of antibiotics on the day of surgery. Furthermore, a path involving computed tomography suggests that postoperative deterioration may have led to its later execution. These results demonstrate that clinically meaningful medical order patterns can be effectively extracted and visualized from real-world data. It confirms that the first evaluation criterion, **"visualization of differences in frequent medical order patterns"**, was successfully fulfilled.

2) Display of Statistical Information: Figure 2 illustrates the display when a node is clicked. The tooltip presents elapsed days from the reference date, test names, support values, and related statistical information. For instance, for a Platelet count, the average result was 20.4, the median value was 19.8, the minimum value was 4.5, and the maximum value was 56.8. The full pattern including the selected node is also highlighted, enhancing understanding of structure and divergence points. This confirms that **"frequent instruction pattern extraction with statistical information"** was appropriately implemented.

3) Visualization of Abnormal Value Occurrence Rates: Figure 3 displays abnormal value rates by elapsed day. Bars are color-coded (red: high, blue: low, green: normal), with separate bars for sequences containing or not containing the pattern. This helps identify when abnormalities are most likely to occur in the clinical process.

Figure 4 shows the tooltip display when hovering over a bar. The tooltip includes support for the pattern and abnormal rate information. In this example, abnormalities were detected in the Ht (hematocrit) tests, and the trend is immediately visible.

This result supports the identification of improvement points in the process and confirms that the **"distribution of a test's normal and abnormal values"** is successfully visualized. The visualization tool serves not only as a display, but also as a powerful aid for clinicians in interpreting data and improving real-world practice.

V. Conclusion and Future Work

A. Conclusion

In this study, we propose a method for analyzing EMRs collected from multiple health care institutions using frequent pattern mining to extract and analyze frequent medical instruction patterns. Unlike previous studies, which mainly focused on single-institution data or specific diseases and lacked statistical analysis, our approach integrates data across multiple institutions and diseases, enabling a more comprehensive analysis. Furthermore, by preserving sequence identifiers, we were able to identify entire sequences corresponding to frequent instruction patterns, allowing for detailed examination.

By linking the test results to frequent instruction patterns and visualizing differences that include abnormal values, our method made it possible to identify not only standard clinical pathways but also critical branching points and exceptional processes. This approach is expected to promote the standardization of clinical workflows and support flexible decision-making tailored to individual patient needs.

B. Future Work

A key direction for future work is to identify the factors that cause branching and variants in frequent instruction sequences. By analyzing the influence of patient attributes such as age, disease history, the distribution of a test's abnormal values distribution, as well as institutional differences in clinical processes, we expect to uncover insights that will further support the standardization and optimization of clinical workflows.

In addition, it is essential to share the results of this research with medical professionals and incorporate their feedback to develop a more practical general-purpose system. Ultimately, the goal is to implement the proposed system as a clinical support tool that contributes to improving the quality and efficiency of health care delivery.

Acknowledgement

This work was partially supported by a Grant-in-Aid for Scientific Research (B) (#24K02943) from the Japan Society for the Promotion of Science.

REFERENCES

[1] H. Yoshihara, "Millennial Medical Record Project: Toward Establishment of Authentic Japanese Version EHR and Secondary Use of Medical Data," *Journal of Information Processing and Management*, vol. 60, no. 11, pp. 767–778, 2018.

[2] S. Arikat and M. Saboor, "Evolving Role of Clinical Laboratories in Precision Medicine: A Narrative Review," *Journal of Laboratory and Precision Medicine*, vol. 9, p. 17, 2024. [Online]. Available: https://jlpm.amegroups.org/article/view/8643

[3] Y. Honda, M. Kushima, T. Yamazaki, K. Araki, and H. Yokota, "Detection and Visualization of Variants in Typical Medical Treatment Sequences," in *Proc. the 3rd VLDB Workshop on Data Management and Analytics for Medicine and Healthcare*, ser. DMAH'17. Springer, 2017, pp. 88–101.

[4] H. H. Le, Y. Horino, T. Yamazaki, K. Araki, and H. Yokota, "Sequential Pattern Mining of Large Combinable Items with Values for a Set-of-items Recommendation," in *Proc. the 34th IEEE Symposium on Computer-based Medical Systems*, ser. CBMS'21. IEEE, 2021, pp. 56–61.

[5] Z. Zhao, Y. Yasumitsu, H. H. Le, T. Yamazaki, K. Araki, and H. Yokota, "Analysis of Transitions in Differences between Frequent Medical-order Sequences for COVID-19," in *Proc. the 36th International Symposium on Computer-Based Medical Systems*. IEEE, 2023, pp. 666–671.

[6] H. H. Le, T. Yamada, Y. Honda, T. Sakamoto, R. Matsuo, T. Yamazaki, K. Araki, and H. Yokota, "Methods for Analyzing Medical-order Sequence Variants in Sequential Pattern Mining for Electronic Medical Record Systems," *ACM Transaction on Computing for Healthcare*, vol. 4, no. 1, pp. 3:1–28, 2023.

[7] H. H. Le, T. Yamada, Y. Honda, M. Kayahara, M. Kushima, K. Araki, and H. Yokota, "Analyzing Sequence Pattern Variants in Sequential Pattern Mining and Its Application to Electronic Medical Record Systems," in *Proc. the 30th International Conference on Database and Expert Systems Applications*, ser. DEXA'19. Springer, 2019, pp. 393–408.

[8] H. H. Le, Y. Yasumitsu, R. Matsuo, T. Yamazaki, and H. Yokota, "A Clustering-based Sequence Variants Analysis Method for Electronic Medical Records of Multimedical Institutions," in *Proc. the seventh IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2024, pp. 653–659.

[9] K. Uragaki, T. Hosaka, Y. Arahori, M. Kushima, T. Yamazaki, K. Araki, and H. Yokota, "Sequential Pattern Mining on Electronic Medical Records with Handling Time Intervals and the Efficacy of Medicines," in *Proc. 2016 IEEE Symposium on Computers and Communication (ISCC)*. IEEE, 2016, pp. 20–25.

[10] A. Rakesh and S. Ramakrishnan, "Mining Sequential Patterns," in *Proc. the 11th International Conference on Data Engineering*, ser. ICDE'95. IEEE, 1995, pp. 3–14.

[11] H. Jiawei, P. Jian, M.-A. Behzad, P. Helen, C. Qiming, D. Umeshwar, and H. MC, "Prefixspan: Mining Sequential Patterns Efficiently by Prefix-projected Pattern Growth," in *Proc. the 17th International Conference on Data Engineering (ICDE)*, 2001, pp. 215–224.

[12] R. V. Purushothama and V. G. Saradhi, "Mining Closed Sequential Patterns in Large Sequence Databases," *International Journal of Database Management Systems*, vol. 7, no. 1, pp. 29–39, 2015.

[13] P. Tzvetkov, X. Yan, and J. Han, "TSP: Mining Top-k Closed Sequential Patterns," *Knowledge and Information Systems*, vol. 7, no. 4, pp. 438–457, 2005.

[14] P. Fournier-Viger, A. Gomariz, T. Gueniche, E. Mwamikazi, and R. Thomas, "TKS: Efficient Mining of Top-k Sequential Patterns," in *Proc. the Ninth International Conference on Advanced Data Mining and Applications*. Springer, 2013, pp. 109–120.

[15] F. Petitjean, T. Li, N. Tatti, and G. I. Webb, "Skopus: Mining Top-k Sequential Patterns Under Leverage," *Data Mining and Knowledge Discovery*, vol. 30, no. 5, pp. 1086–1111, 2016.

[16] T.-T. Pham, T. Do, A. Nguyen, B. Vo, and T.-P. Hong, "An Efficient Method for Mining Top-k Closed Sequential Patterns," *IEEE Access*, vol. 8, pp. 118 156–118 163, 2020.

[17] D.-A. Chiang, Y.-F. Wang, S.-L. Lee, and C.-J. Lin, "Goal-oriented Sequential Pattern for Network banking Churn Analysis," *Expert systems with applications*, vol. 25, no. 3, pp. 293–302, 2003.

[18] C. Chand, A. Thakkar, and A. Ganatra, "Target Oriented Sequential Pattern Mining using Recency and Monetary Constraints," *International Journal of Computer Applications*, vol. 45, no. 10, 2012.

[19] G. Huang, W. Gan, and P. S. Yu, "TaSPM: Targeted Sequential Pattern Mining," *arXiv preprint arXiv:2202.13202*, 2022.

[20] C. Yen-Liang, C. Mei-Ching, and K. Ming-Tat, "Discovering Time-interval Sequential Patterns in Sequence Databases," *Expert Systems with Applications*, vol. 25, no. 3, pp. 343–354, 2003.

[21] H. H. Le, H. Edman, Y. Honda, M. Kushima, T. Yamazaki, K. Araki, and H. Yokota, "Fast Generation of Clinical Pathways Including Time Intervals in Sequential Pattern Mining on Electronic Medical Record Systems," in *Proc. the 4th International Conference on Computer Science and Computational Intelligent*, ser. CSCI'21, 2017, pp. 1726–1731.

[22] Y. Li, H. H. Le, R. Matsuo, T. Yamazaki, K. Araki, and H. Yokota, "Comparison of sequence variants and the application in electronic medical records," in *Proc. the 33rd Database and Expert Systems Applications, Part 2*, ser. DEXA'22, 2022, pp. 117–130.

[23] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, "PrefixSpan: Mining Sequential Patterns by Prefix-projected Growth," in *Proceedings of the 17th International Conference on Data Engineering*, 2001, pp. 215–224.