Extracting and Visualizing Frequent Medical Instruction Patterns with Statistical Insights from Multi-Institutional Electronic Medical Record Data

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*—Despite the widespread adoption of electronic medical records (EMR), the variations in format and terminology across institutions hinder inter-institutional comparisons and feature extraction. This paper proposes a demonstration of extracting frequent disease-specific instruction sequences and efficiently visualizing them with statistical insights, e.g. statistical trends, and abnormal inspection result rates from real multi-institutional EMR data. The utility of the developed visualization tool was presented for decision support and improving clinical processes.**

***Index Terms*—electronic medical record, pattern and association rule mining, visualization**

I. Introduction

The widespread adoption of electronic medical records (EMR) has facilitated aggregation and secondary use, thereby promoting the standardization of clinical processes. However, challenges remain in utilizing EMR data from multiple medical institutions and supporting personalized medical care [1]. Existing studies have tended to focus on single diseases or institutions when analyzing the factors that influence medical instruction [2], making insufficient cross-disease and inter-institutional comparisons. Furthermore, the extraction of statistical information has been inadequate, and its relationship with medical instruction patterns has not been visualized effectively, which limits its use in clinical decision making. In detail, while existing visualization tools depict the branching of frequent medical instruction patterns, they do not incorporate the underlying statistical information or abnormal inspection values.

This paper proposes a demonstration of a practical analysis and visualization of frequent medical instruction patterns that deliver useful insights, such as statistical information, and abnormal inspection result rates, thereby enhancing the utilization of medical data.

II. Proposed Method

A. Construction of Medical Instruction Sequences

The proposed method first organizes clinical records from EMR data, constructs medical instruction sequences by calculating elapsed days based on critical medical instructions (e.g., surgery), and assigns a unique patient identifier.

Definition 1 (Sorting Rules for Medical Instructions)**.** *The medical instructions recorded on the same day are sorted in the following order, with the instructions within the same category arranged lexicographically.*

$$\textit{Surgery} \rightarrow \textit{Medication} \rightarrow \textit{Inspection} \rightarrow \textit{Treatment}$$

Definition 2 (Medical Instruction Sequence with Elapsed Days)**.** *The medical instruction sequence for a patient p_i is defined in the following format, incorporating the elapsed days d_j from the reference date t_0 (Day 0):*

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

where $d_j = t_j - t_0$ represents the elapsed days, and a_j denotes the medical instruction.

B. Extraction of Frequent Medical Instruction Patterns

The T-PrefixSpan algorithm [3] extracts frequent medical instruction patterns while preserving time intervals and sequential information. Simultaneously, statistical information on inspection values is extracted and utilized to assess disease progression and treatment effectiveness.

Definition 3 (Frequent Medical Instruction Patterns with Statistical Information)**.** *Let $X(\alpha)$ represent the results for a specific inspection α corresponding to a frequent medical instruction pattern fs. Let x_j ($j = 1, 2, \ldots, k$) be the observed values, where k represents the total number of observations. The following statistical measures are computed:*

$$\mathrm{Mean}(X(\alpha)) = \tfrac{1}{k}\textstyle\sum_{j=1}^{k} x_j \qquad \mathrm{Median}(X(\alpha)) = x_{\lceil \frac{k}{2} \rceil}$$
$$\mathrm{Max}(X(\alpha)) = \max(X(\alpha)) \quad \mathrm{Min}(X(\alpha)) = \min(X(\alpha))$$

C. Calculation of Inspection Result Rates

For each frequent medical instruction pattern fs, the inspection result rates are calculated for patients with the target disease. The abnormal inspection result rate is defined as the proportion of inspection results outside the normal range $[B(\alpha), T(\alpha)]$, to assess abnormal trends and aid diagnosis.

Definition 4 (Abnormal Inspection Result Rate)**.** *First, the set of medical instruction sequences for the target disease is denoted as $D = \{S_{p_1}, S_{p_2}, \ldots, S_{p_m}\}$. The sequences containing the frequent pattern fs are grouped as D_{in}, while those that do not contain fs are grouped as D_{out}:*

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

The high abnormal rate (HR), low abnormal rate (LR), and normal rate (NR) for each elapsed day d_j are defined as follows. Here, $X(\alpha, d_j)$ is the set of results for a specific inspection α obtained on day d_j:

$$HR(\alpha, d_j) = \frac{|x_i \in X(\alpha, d_j) \mid x_i > T(\alpha)|}{|X(\alpha, d_j)|}$$
$$LR(\alpha, d_j) = \frac{|x_i \in X(\alpha, d_j) \mid x_i < B(\alpha)|}{|X(\alpha, d_j)|}$$
$$NR(\alpha, d_j) = 1 - (HR(\alpha, d_j) + LR(\alpha, d_j))$$

D. Visualization

This study visualizes frequent medical instruction patterns and abnormal inspection result rates to clarify characteristics in clinical processes. Frequent medical instruction patterns are displayed in a time-sequenced layout, with nodes positioned based on support values for intuitive interpretation. Selecting a node reveals detailed statistical information. Abnormal inspection result rates are represented as bar charts, visualizing the proportions of "High", "Normal", and "Low" values. Filtering by inspection items and viewing detailed statistics are also supported, enhancing the efficiency of pattern analysis.

III. Evaluation Experiment

A. Experimental Environment

In this experiment, the PrefixSpan library[1] was modified to implement functions equivalent to T-PrefixSpan to extract frequent patterns. The dataset used in this experiment contains 59,598 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).

B. Experimental Results

As an example, Fig. 1 visualizes frequent medical instruction patterns for 9,962 patients with malignant lung tumors[2]. The patterns were extracted with a minimum support threshold (MinSup) of 40%. 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 indicate that clinically valid medical instruction patterns were successfully extracted from real-world data.

Fig. 2 visualizes the distribution of inspection result rates for patients based on the medical instruction patterns in Fig. 1. Clicking a node in Fig. 1 separates patients into those who follow the selected path (frequent patterns) and those who do not (non-frequent patterns), and their abnormal inspection rates are then displayed. The bars represent high, low, and normal values in red, blue, and green, respectively. Hovering over a bar

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

[2]The demo of the proposals with all functionalities for other diseases can be presented when accepted.

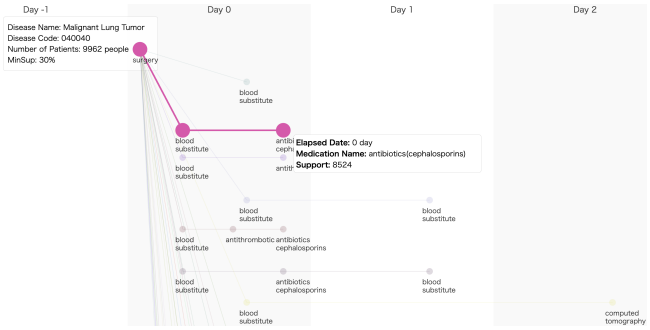

Fig. 1. Frequent medical instruction patterns

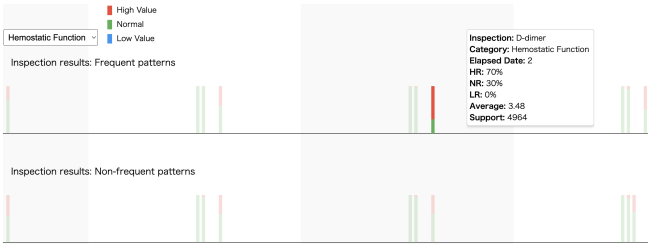

Fig. 2. Visualization of the distribution of inspection result rates

displays details, such as the inspection name, abnormality ratio, and support value. On day 2, the D-dimer test in the frequent pattern group shows a high value rate of 70% and a low value rate of 0%, whereas in the non-frequent pattern group, the high value rate is 39% and the low value rate remains 0%. This highlights the differences in abnormal result distributions between the frequent and non-frequent instruction sequences.

The proposed method intuitively identifies patient characteristics associated with specific medical instruction patterns.

IV. Conclusion

This study proposed a method for extracting and visualizing frequent medical instruction sequences from EMR data across multiple medical institutions delivering valuable statistical insights, such as statistical information and the distribution of inspection values. The present findings could be expected to help medical staff grasp the differences between frequent and remaining sequences, which would contribute to the understanding of clinical processes and promote individualized treatment planning.

Acknowledgments

This research was partially supported by a grant from the Japan Society for the Promotion of Science (#24K02943).

References

[1] J. Larry Jameson and Dan L. Longo. Personalized medicine: A reality in the oncology clinic. *Journal of Clinical Oncology*, 36(6):536–544, 2018.

[2] H.H. Le et al. Methods for analyzing medical-order sequence variants in sequential pattern mining for electronic medical record systems. *ACM Transactions on Computing for Healthcare*, 4(1), March 2023.

[3] K. Uragaki et al. Sequential pattern mining on electronic medical records with handling time intervals and the efficacy of medicines. In *Proceeding of the 21st IEEE International Symposium on Computers and Communications*, pages 20–25, 2016.