Trend Event Pattern Mining*

Abdul Hadi and Talha Farhat

CureMD Research, 80 Pine St 21st Floor, New York, NY 10005, United States
{talha.farhat,hadi.khadim}@curemd.com
https://www.curemd.com

Abstract. The surge in Electronic Health Records (EHR) utilization provides a unique opportunity to enhance patient care through advanced data analysis techniques. This study leverages trend-event pattern mining within EHRs to offer significant insights into patient responses to medical treatments over time, focusing particularly on chronic conditions. By applying temporal data mining, we aim to identify and categorize key patterns associated with treatment events, enhancing the understanding and prediction of patient outcomes. The study outlines a framework that incorporates a sophisticated data preprocessing pipeline, innovative temporal and value abstraction methods, and a dynamic scoring system to assess the impact of treatment events on patient health. Our results indicate that close temporal associations between treatment events and physiological readings yield the most clinically relevant insights, thereby suggesting modifications to current treatment strategies and helping tailor interventions to individual patient needs. Overall, this research not only underscores the potential of using temporal pattern mining in healthcare to improve patient outcomes but also sets the stage for further integration of these techniques into clinical practice, particularly for managing chronic diseases.

Keywords: Temporal Patterns · Electronic Health Records (EHR) · Unsupervised Learning · Healthcare Prognosis · Clinical Data Analysis.

1 Introduction

The increasing availability of Electronic Health Records (EHR) presents unprecedented opportunities for advancing medical research and enhancing patient care. EHRs offer a comprehensive view of patient histories, treatments, and outcomes, yet they also pose significant challenges due to their voluminous and noisy nature. The effective analysis of EHR data requires sophisticated data mining techniques that can transform raw data into actionable medical insights.

1.1 Research Context

The utilization of temporal data mining in healthcare is a rapidly evolving area of research that addresses the need for advanced analytical tools capable of

* Supported by CureMD.

interpreting the complex dynamics of patient data over time. This approach is particularly relevant in the context of chronic diseases, where understanding the long-term patterns of patient responses to treatments can lead to more effective management strategies and improved patient outcomes.

1.2 Significance of the Study

This study is positioned at the intersection of data science and healthcare, aiming to harness the power of trend-event pattern mining to predict and evaluate the efficacy of medical treatments. By focusing on the temporal aspects of EHR data, this research seeks to uncover patterns that are not readily apparent through traditional analysis methods. The ability to predict patient outcomes following specific treatment events has significant implications for personalized medicine, offering a pathway to tailor medical interventions to individual patient needs based on historical data.

1.3 Objectives

The primary objective of this research is to develop and validate a trend-event pattern mining framework that:

- Identifies and categorizes temporal patterns in patient data that correlate with key treatment events.
- Evaluates the impact of these patterns on patient outcomes to inform treatment decisions.
- Provides a scalable and reproducible methodology that can be applied across various medical conditions and treatment modalities.

1.4 Structure of the Report

This report is organized into several sections that chronologically detail the research process:

1. **Related Work:** Discusses previous studies and frameworks relevant to temporal data mining in healthcare.
2. **Methodology:** Describes the data processing, pattern extraction, and analysis techniques employed in this study.
3. **Results:** Presents the findings from the application of the trend-event pattern mining framework.
4. **Discussion:** Interprets the results in the context of existing knowledge and potential implications for clinical practice.
5. **Conclusion and Future Work:** Summarizes the study's contributions and outlines directions for future research.

This introduction sets the stage for a detailed exploration of how trend-event pattern mining can transform the landscape of healthcare analytics, highlighting the innovative aspects of this research and its potential to contribute significantly to the field of personalized medicine.

2 Related Work

The integration of temporal data mining techniques in healthcare has become a pivotal tool for understanding and predicting patient outcomes. However, the inherently noisy nature of Electronic Health Records (EHR) data presents challenges that have historically limited the implementation of purely knowledge-based data mining algorithms. With the proliferation of machine learning techniques, traditional knowledge-based algorithms have often been relegated to a secondary role, resulting in sparse literature on EHR data mining.

Notably, the work of Batal et al. (2012) marks a significant advancement in this field. They introduced the concept of recent temporal pattern mining, focusing on the detection of events in multivariate time series by identifying frequent patterns that are temporally proximal to the event of interest. Their approach emphasized the utility of temporal abstractions to simplify time series data, thereby enhancing the extraction of actionable insights and improving the predictability of acute medical events.

Building on this foundation, Batal et al. (2013) further refined their methodologies by introducing the Minimal Predictive Temporal Patterns framework. This framework aims to simplify the mined patterns and concentrate on those that are most predictive of clinical outcomes. By focusing on the significance of the patterns in terms of their predictive power, this method supports more precise patient monitoring and diagnosis.

Moreover, Mantovani et al. (2019, 2021) expanded the scope of this research by exploring trend-event patterns (TE-Ps). They hypothesized that identifying trends preceding and following a medical event, such as medication administration, could be instrumental in predicting the effectiveness of a treatment. Their studies, conducted on patients in Intensive Care Units, focused on mining patterns from vital signs like mean arterial pressure in patients with sepsis. They developed an algorithm to validate each identified trend by counting the occurrences of readings around the medication event, imposing constraints to ensure only the most prominent patterns were selected. Although their approach provided valuable insights into pattern mining strategies, the specific nature of their data differed significantly from the EHR data typically encountered in our healthcare data setting.

These seminal studies have provided a robust framework and set a precedent for the application of data mining techniques in the analysis of temporal EHR data, informing the methodology and approach of our research.

3 Methodology

This section will first focus on the concept behind the mining algorithm used and then

3.1 ETL Process

To manage the extensive datasets of Electronic Health Records (EHR), we implemented a robust ETL (Extraction, Transformation, and Loading) process. Initially, attempts to load comprehensive patient data in a single batch led to memory overflow errors. To mitigate this, data was extracted in manageable chunks using the Python library `pandas`. This approach not only resolved memory issues but also optimized data handling efficiency. Each chunk was transformed to ensure data format consistency and then loaded into a CSV format for persistent storage. This method significantly reduced the need for continuous server queries, enhancing the overall data processing speed and reliability.

3.2 Data Cleaning and Standardization

Critical to the success of any data-driven project is the integrity and accuracy of the data used. In this project, several steps were undertaken to clean and standardize the EHR data:

- **BMI Standardization:** Discrepancies in BMI calculations were addressed by converting all height and weight measurements to the International System of Units (SI). This standardization was crucial where data entry errors had occurred, such as mixing measurements in pounds and inches with those in kilograms and meters.
- **Mean Arterial Pressure (MAP) Adjustment:** MAP readings, a vital indicator of cardiovascular health, were standardized by applying a statistical method to clip outliers based on the inter-quartile range (IQR). This approach, validated through medical consultation, ensured that the data used in subsequent analyses were within a clinically acceptable range.
- **Hypertension: Units Standardization** There wa a lot of variation in the units being used for the Hypertension Lab test. Fig 1 shows the distribution of the tests along with their units. These were cleaned by grouping similar tests along with simplified units as depicted in 2 .

3.3 Tuple Extraction

Tuple extraction was designed to create actionable sequences of data that reflect patient conditions over time relative to medication administration:

- **Sequence Construction:** Using the cleaned data, the most recent and preceding lab and vital readings for each patient were identified. Medications administered between these two readings were then linked to form a tuple representing a temporal sequence.
- **Tuple Format:** Each tuple was formatted as follows:
  (patient_ID, measured_metric_name, Days Before, Value_before_medication, Med_GPI_Code, Days After, Value_after_medication). This structure facilitated the subsequent analysis of treatment effects over specified time intervals.

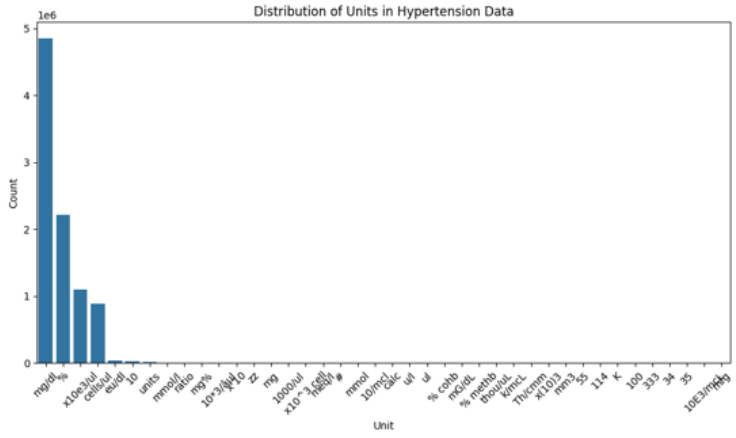

Fig. 1: No. of units available in the uncleaned Hypertension Lab results.

3.4 Abstraction

Given the inherent noisiness and inconsistency in EHR data, abstraction was employed to distill the raw data into more analytically useful forms:

- **Value Categorization:** This involved dividing continuous measurements into discrete categories based on clinically relevant thresholds, as verified by domain experts. This categorization helped to standardize the analysis by reducing variability due to minor measurement deviations.
- **Temporal Abstraction:** Recognizing the importance of timing in medical interventions, this abstraction focused on the significance of data points relative to their proximity to medication events. The hypothesis that readings closer to medication events are more indicative of immediate treatment effects guided this analysis.

3.5 Temporal Abstraction

Temporal abstraction is essential for distilling complex EHR time series data into interpretable and clinically relevant patterns. This approach quantifies the temporal proximity of observations relative to medication events, which significantly influences their clinical relevance.

Two-Phase Exponential Decay To model the validity of lab test results over time, we employ a two-phase exponential decay function, represented as $D(t)$. This function determines the penalty applied to data points based on their temporal distance from a critical event, such as medication administration. The decay function is defined as follows:

$$D(t) = \begin{cases} e^{-r_1 \cdot t} & \text{for } t \leq T \\ e^{-r_2 \cdot (t-T)} & \text{for } t > T \end{cases}$$

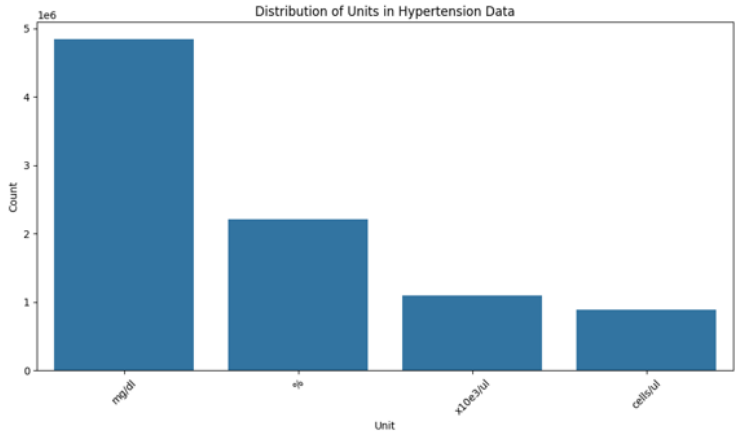

Fig. 2: Simplified units in hypertension lab data after cleaning

where t is the time elapsed from the event, T is the threshold time distinguishing the two phases of decay, r_1 and r_2 are the decay rates for each phase, respectively. The first phase ($t \leq T$) represents a slower rate of decay, suitable for the immediate post-medication period where changes are most crucial. The second phase ($t > T$) applies a faster decay, reflecting the diminishing relevance of data points as they move further from the event time.

This dual-phase model allows for a more nuanced analysis of data, recognizing that the importance of physiological readings decreases at different rates as time progresses from a treatment event. By applying this decay function, we can prioritize recent data points without dismissing the potential insights from older data that may still hold clinical value.

Scoring Mechanism In addition to the two-phase exponential decay, we introduce a scoring mechanism that quantifies the significance of each tuple based on its temporal proximity to a treatment event. This scoring formula is designed to highlight tuples where both pre-event and post-event readings are closely aligned with the event, as these are presumed to provide the most clinically relevant insights.

The scoring for each tuple is computed using the following formula:

$$\text{Score} = \text{Penalty for Days Before} \times \text{Penalty for Days After}$$

where:

- **Penalty for Days Before** is calculated as $e^{-r_1 \cdot t_1}$ for $t_1 \leq T$ or $e^{-r_2 \cdot (t_1 - T)}$ for $t_1 > T$, where t_1 is the time before the medication event.
- **Penalty for Days After** follows a similar decay function, calculated for t_2, the time after the medication event.

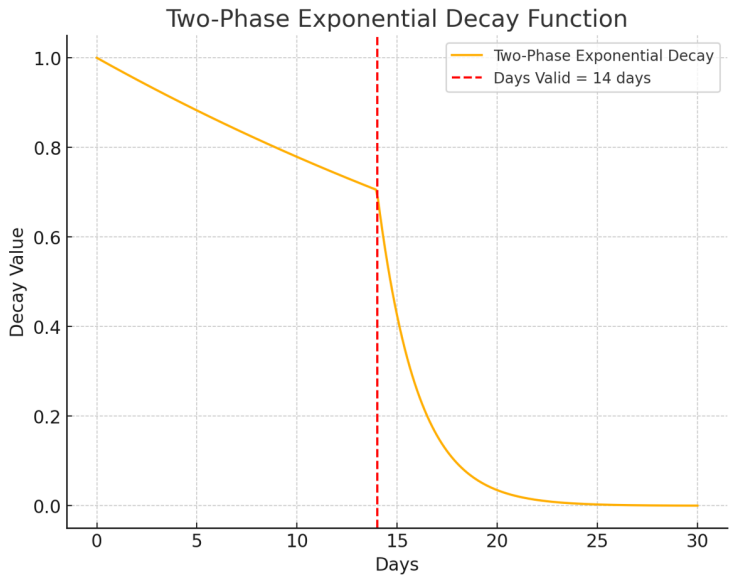

Fig. 3: Visual Representation of the two phased Decay. The initial decay to the no. of valid days for a test is at a lower rate. After the test validity is over, the decay is much faster

This scoring model ensures that the highest scores are assigned to tuples where both readings — before and after the medication event — occur within a narrow time window. Such a configuration implies a strong temporal correlation between the medication event and the observed changes in the patient's metrics, thereby maximizing the clinical relevance of the tuple. By incorporating this scoring mechanism, we can effectively prioritize data points that are most likely to yield meaningful insights into the treatment's impact.

3.6 Grouping Based on Trends

After abstraction, the data were grouped to identify patterns in treatment outcomes:

- **Aggregation Strategy:** Data points were aggregated by GPI codes and categorized before and after medication events. A penalty system was introduced where data points further from the medication event received decreasing weights, reflecting their reduced impact on immediate treatment outcomes.
- **Statistical Support:** The aggregation included summing penalty scores and counting occurrences within each group to provide quantitative support for observed trends.

Aggregation and Normalization of Scores After calculating the scores for individual tuples, we aggregate these scores to analyze broader trends in the data. The aggregation process involves summing all scores that fall under the same grouping, defined by specific pre-event and post-event categories.

Additionally, to enhance the interpretability of these scores:

- A new column, *Count*, is added to the dataset, which records the total number of occurrences for each grouping. This count serves as a quantitative measure of the support for the observed scores within each category.
- The aggregated scores are then normalized to account for variations in the frequency of different categories. The normalization is performed using the following formula:

$$\text{Normalized Score} = \frac{\text{Score}(\text{Cat\_After}|\text{Cat\_Before})}{\sum \text{Score}(\text{Cat\_Before})}$$

  where $\text{Score}(\text{Cat\_After}|\text{Cat\_Before})$ represents the aggregated score for tuples transitioning from a specific pre-event category to a post-event category, and $\sum \text{Score}(\text{Cat\_Before})$ is the sum of all scores in the pre-event category across all groupings.

This approach ensures that the results are not only reflective of the raw score values but are also adjusted to reflect the relative frequency of occurrence of each category. Such normalization allows for more

3.7 Explainability

To ensure that the results of our complex data processing were interpretable:

- **Transparency in Data Processing:** A system was implemented to allow healthcare professionals to view the detailed patient data underlying each aggregated score. This functionality is crucial for validating the automated insights and supports clinical decision-making by providing a traceable path back to the foundational data. Fig 4 shows the graph that relates the scores of each trend

4 Results

4.1 Data Analysis Outcomes

The application of the trend-event pattern mining framework to Electronic Health Records (EHR) data yielded significant insights into patient treatment patterns:

- **Patient Cohort Analysis:** Of the 450,000 patients analyzed, focusing on those diagnosed with hypertension, a total of 3.3 million valid data tuples were extracted. This robust dataset allowed for a comprehensive analysis of treatment efficacy.

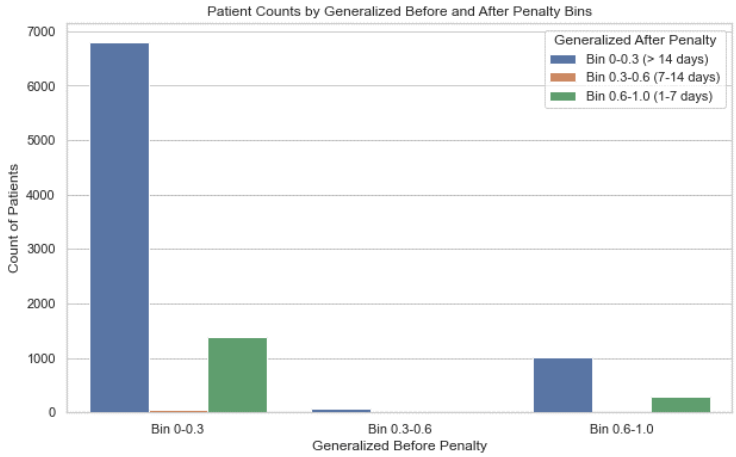

Fig. 4: Detailed View of the reasoning behind the aggregated Score assigned to each Extracted Trend. The green bars depict results that are conducted near to the medication

– **Treatment Efficacy:** For hypertension patients, the analysis identified a clear pattern in the effectiveness of the drug with GPI code 39400010. This medication moved mean arterial pressure readings from high to normal ranges in over 5,000 cases, indicating a high level of effectiveness.

4.2 Grouping and Trend Identification

The grouping based on trends provided a novel insight into the categorization of patient responses to treatments:

– **Statistical Significance:** The grouped data, aggregated by GPI code and pre/post-medication categories, showed statistically significant trends in medication outcomes. This was supported by the aggregation of penalty scores, where a total of 22,000 groupings were analyzed to determine the most effective treatment categories.
– **Support Counts:** Each grouping was quantified with a count of occurrences, providing robust support for the observed trends. For example, medications classified within the GPI code 3400, belonging to calcium blockers, showed significant efficacy in altering blood vessel calcium dynamics, verified across numerous patient records. Fig5 shows the Detailed trend results against Calcium Blocker Medicine for Hypertension Patients having **High** categorized Mean Arterial Pressure.

4.3 Implications for Future Research

The insights gained from this study pave the way for extended applications of the trend-event pattern mining framework:

Vital Name	Category Before	Category After	GPI Code	Count	Score
mean_arterial_pressure	H	H	3400010	5292	0.491
mean_arterial_pressure	H	N	3400010	5353	0.483
mean_arterial_pressure	H	VH	3400010	542	0.026
mean_arterial_pressure	H	L	3400010	20	0.0
mean_arterial_pressure	H	VL	3400010	3	0.0

Fig. 5: Detailed trend results against Calcium Blocker Medicine for Hypertension Patients Having High Mean Arterial Pressure

- **Broader Disease Applications:** The successful application to hypertension suggests potential extensions to other common diseases, using the established methodology to analyze different sets of vitals and medication events.
- **Integration into Clinical Practice:** The framework offers a potential for integration into clinical decision-making tools, providing a data-driven basis for personalizing treatment plans.

4.4 Conclusion

The results of this study demonstrate the efficacy of the trend-event pattern mining framework in extracting meaningful patterns from large-scale EHR data. The findings not only reinforce the validity of the methodologies employed but also offer promising avenues for enhancing personalized medicine practices. Future work will focus on expanding the scope of diseases analyzed and integrating insights into practical clinical workflows.

5 References

1. Batal, I., Valizadegan, H., Cooper, G. F., & Hauskrecht, M. (2012). A pattern mining approach for classifying multivariate temporal data. *Bioinformatics*, 28(18), i233-i241. doi:10.1093/bioinformatics/bts376
2. Batal, I., Valizadegan, H., Cooper, G. F., & Hauskrecht, M. (2013). Mining recent temporal patterns for event detection in multivariate time series data. *Knowledge and Information Systems*, 40(3), 611-635. doi:10.1007/s10115-012-0585-9
3. M. Mantovani, C. Combi and M. Zeggiotti, "Discovering and Analyzing Trend-Event Patterns on Clinical Data," 2019 IEEE International Conference on Healthcare Informatics (ICHI), Xi'an, China, 2019, pp. 1-10, doi: 10.1109/ICHI.2019.8904774.
4. Mantovani, R. G., Rossi, A. L. D., Vanschoren, J., Bischl, B., & de Carvalho, A. C. P. L. F. (2021). Effective trend-event pattern mining in intensive care patient data for proactive monitoring. *Journal of Biomedical Informatics*, 114, 103637. doi:10.1016/j.jbi.2020.103637