Analyzing Time Attributes in Temporal Event Sequences

Jessica Magallanes*
Department of Computer Science
INSIGNEO Institute for *in silico* Medicine
University of Sheffield
Sheffield, United Kingdom

Lindsey van Gemeren
Sheffield Teaching Hospitals NHS Foundation Trust
INSIGNEO Institute for *in silico* Medicine
University of Sheffield
Sheffield, United Kingdom

Steven Wood
Sheffield Teaching Hospitals NHS Foundation Trust
INSIGNEO Institute for *in silico* Medicine
University of Sheffield
Sheffield, United Kingdom

Maria-Cruz Villa-Uriol†
Department of Computer Science
INSIGNEO Institute for *in silico* Medicine
University of Sheffield
Sheffield, United Kingdom

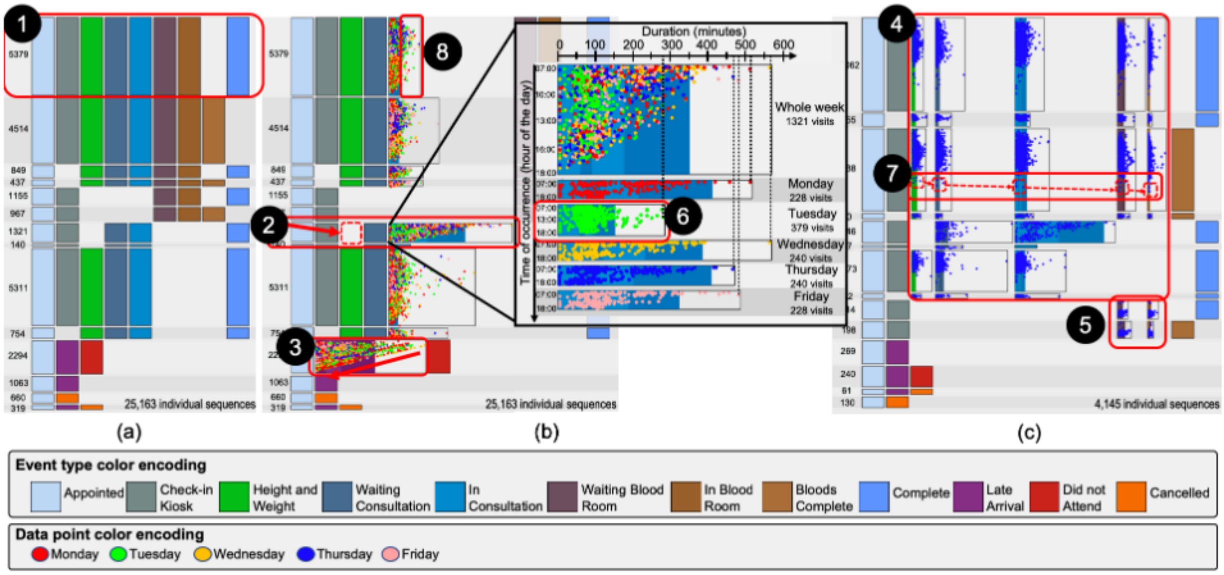


Figure 1: Three different configurations for the Sequential and Time Patterns overview using the Rheumatology dataset. (a) Sequential and Time Patterns overview where all events are encoded as point events. (b) The events "In Consultation" and "Late Arrival" are expanded for analysis. The expanded events show the proposed event encoding for time attributes, which allow to identify trends and outliers with respect to duration and time of occurrence. For a selected sequence, the event "In Consultation" is broken down into days of the week for further exploration. (c) The Sequential and Time patterns overview is filtered to show only the events happening on Thursdays. The highlighted numbers represent findings which are explained in the taxonomy of findings proposed.

Abstract

Event data is present in a variety of domains such as electronic health records, daily living activities and web clickstream records. Current visualization methods to explore event data focus on discovering sequential patterns but present limitations when studying time attributes in event sequences. Time attributes are especially important when studying waiting times or lengths of visit in patient flow analysis. We propose a visual analytics methodology that allows the identification of trends and outliers in respect of duration and time of occurrence in event sequences. The proposed method presents event data using a single Sequential and Time Patterns overview. User-driven alignment by multiple events, sorting by sequence similarity and a novel visual encoding of events allows the comparison of time trends across and within sequences. The proposed visualization allows the derivation of findings that otherwise could not be obtained using traditional visualizations. The proposed methodology has been applied to a real-world dataset provided by Sheffield Teaching Hospitals NHS Foundation Trust, for which four classes of conclusions were derived.

Index Terms: Human-centered computing—Visualization—Visualization techniques; Human-centered computing—Visual analytics; Applied computing—Health care information systems

1 Introduction

Event logs are routinely recorded in a variety of domains such as electronic health records, daily living activities and web clickstream records. The analysis of event data can provide valuable insights into processes and the behaviour of individuals.

Existing visualization techniques focus on the visual encoding of sequential patterns, but present limitations when representing time attributes. Firstly, the time of occurrence is not explicitly encoded in the overview of sequences. Secondly, the duration of events (if visualized) is encoded using the average value. Exclusively representing average duration means that individuals with unusual durations (*outliers*) are easily missed and, therefore, overlooked. The exact time at which individual events occur might not be of interest alone, but when combined with other variables, trends in time can be obtained to provide deeper insights. For example, visualizing time attributes is essential in the context of health care delivery. Waiting times and lengths of stay are key performance indicators required

*e-mail: jgmagallanescastaneda1@sheffield.ac.uk
†e-mail: m.villa-uriol@sheffield.ac.uk

to optimize the use of resources whilst maximizing the quality of patient care.

We propose a generic methodology which is able to integrate the extraction and interactive visualization of sequential and time patterns from temporal event data. The identification of *time patterns* is facilitated by a novel visual encoding capable of representing duration and time of occurrence *at event level*. *At sequence level*, a set of the most relevant unique sequences is obtained first, then sorted by sequence similarity, and finally visually aligned to obtain *sequential patterns*.

The proposed methodology allows users to interact with the data and focus their analysis on common patterns or on anomalous situations that might be of interest. We have applied the methodology to a real-world dataset in the healthcare domain, and present a taxonomy of findings that can be generalized to other application domains. The contributions of the present work are:

- **Visual encoding of time attributes at event level:** a novel visual encoding which aggregates the duration and time of occurrence of events.
- **Integration of time and sequential patterns into a single overview:** our approach allows the analysis of patterns that combine time attributes and the sequential order of event sequences.

2 Related Work

2.1 Visual analytics of Temporal Event Sequences

A variety of visual analytic techniques for event data have been proposed. Typically, an overview of the main sequential patterns is provided as a tree-like view, encoding sequences by frequency and average duration [16, 22]. Other existing views include: icicle plots [13, 16], sankey diagrams [18, 22], transition matrices [23], state transitions graphs [11, 20], and list of glyphs [5]. Various techniques have addressed the identification of sequential patterns either by using frequent sequence mining (FSM) [10, 14, 18] or clustering methods [5, 8, 22].

The overview of sequential patterns can be manipulated using a wide range of operations which are application-dependent.

In general, these operations can be grouped into three categories: transformations, queries and alignment. Transformations allow the simplification of the data overview [16], for example either by merging similar event types into a single one or by removing records that are not of interest. Queries allow the filtering of data being visualized according to a set of events or temporal constraints of variable complexity. Existing strategies include: visual queries [17], regular expressions [4] or milestone events [7, 13]. Alignment of event sequences by a selected event, target the exploration of the subset of events happening right before and after the alignment point. Typically, only one event can be selected as alignment point [5, 16, 22], although more recently Eventpad [4] implemented a Multiple Sequence Alignment (MSA) strategy in which the user is able to modify the gap cost in the algorithm proposed by Bose et al. [3].

2.2 Overview of time attributes

Existing visual analytics methods mainly focus on visualizing the sequential order of the events [16, 18, 22]. Currently, the visual encoding of time attributes is limited. Time of occurrence is always implicit in the sequential ordering of the sequences. However, no explicit time attributes (e.g. 3pm, Monday, May, 2019) are fully visually encoded in the overview.

Generally, a separate secondary view is required to review the time of occurrence for a selected record. ActiviTree [20] visualizes the distribution of sequences across the time of the day, using a secondary view. In LifeLines2 [21], the distribution of the frequency of selected records through time is visualized using a histogram (i.e. the number of occasions a particular event happens on a specific date). Events within a time range before and after an alignment point can also be analyzed. However, this method does not aggregate sequences, and frequency distribution is only shown for the selected event.

TimeSpan [15] uses stacked bar charts to indicate the duration of events related to a stroke treatment process, a line chart is used to study trends in duration through monthly intervals. However, they assume that sequences do not vary in the ordering of events, meaning sequential patterns are not included.

In methods where the duration of events is visually encoded, the width of an event is scaled proportionally to the average duration. This approach ignores the distribution of the duration and the presence of outliers. Duration outliers can be defined as observations with a duration which appears to be inconsistent with the remainder of the data [2]. Previous literature [20] indicates the importance of identifying infrequent sequences as outliers, but no emphasis has been made to pinpoint duration outliers.

Some techniques offer the possibility of filtering events by their duration. Eventflow [16] allows querying event sequences after specifying a time window (e.g. displaying only events with a duration above thirty minutes), while Eventpad [4] provides a histogram to inspect event attributes separately. However, this information is not encoded in the sequential patterns overview.

3 Time attributes at event level

Event data refers to a set of time-stamped events, where an *event* can be expressed as what happened (*event type*), when (*time of occurrence*), for how long (*duration*), and who was involved (*identifier*). For each *identifier*, an ordered list of events can be obtained (*event sequence*); and for the complete set of event sequences in a dataset, the *unique sequences* can be extracted by grouping those event sequences sharing the same ordered set of events exactly.

We propose a novel visual encoding to represent time attributes at event level, allowing the identification of trends and outliers with regard to duration and time of occurrence. The proposed approach also allows exploring time of occurrence at multiple time scales, and several levels of detail for the duration of events.

3.1 Visualizing duration and time of occurrence

The proposed visual encoding of events is inspired in boxplots [19]. For univariate data, boxplots have proved excellent to understand a whole sample of observations. They present multiple benefits: highlight of outliers, resistance to extreme values, representation of variability and skewness of the sample.

An *event box* aggregates events of a particular type (see Fig. 2a), its *height* is proportional to the number of records, and its *width* proportional to the maximum variation ($Q_4 - Q_0$) in duration. An event box is composed by *quartiles* and *data points*.

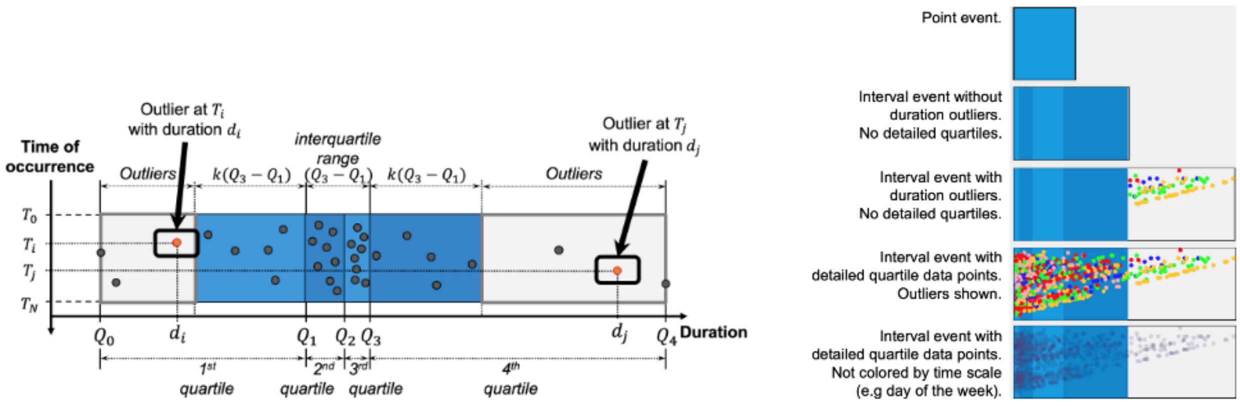
Quartiles are represented as colored sub boxes delimited by a duration value Q_i. In this case, Q_0 is the minimum duration and Q_4 the maximum duration. Q_1,Q_2 and Q_3 correspond to the 25^{th}, 50^{th} and 75^{th} percentile respectively. Each data point in the visualization represents an individual event occurrence. Depending on their duration, data points can be either considered as *outliers* or *quartile points*.

Using Tukey's [6] definition, outliers are identified as data points outside the range:

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)],$$

where $(Q_3 - Q_1)$ is the interquartile range and traditionally $k = 1.5$. Opposite to outliers, quartile points are visually identified as those within the colored sub boxes.

Data points are placed along the *horizontal* axis according to their duration, and placed in the *vertical* axis according to their time of occurrence. The scale of the *vertical* axis is given by the range T_0 to T_N, where T_0 is the minimum and T_N represents the maximum time


Outlier at T_i with duration d_i
Time of occurrence
Outliers
$k(Q_3 - Q_1)$
interquartile range $(Q_3 - Q_1)$
$k(Q_3 - Q_1)$
Outlier at T_j with duration d_j
Outliers
T_0
T_i
T_j
T_N
Q_0
d_i
Q_1
Q_2
Q_3
d_j
Q_4
Duration
1st quartile
2nd quartile
3rd quartile
4th quartile
Point event.
Interval event without duration outliers. No detailed quartiles.
Interval event with duration outliers. No detailed quartiles.
Interval event with detailed quartile data points. Outliers shown.
Interval event with detailed quartile data points. Not colored by time scale (e.g day of the week).

(a) (b)

Figure 2: (a) Proposed visual encoding of an event box which is used to aggregate events. Data points represent individual event occurrences. Quartiles are delimited by duration ranges, where Q_0 is the minimum duration and Q_4 the maximum. Data points are located in the horizontal axis according to duration and in the vertical axis according to time of occurrence. (b) Example of an event box at five different levels of detail. Hiding or showing quartile and outlier data points, and changing the colour of the data points, result in different levels of detail.

of occurrence being visualized. This arrangement of data points allows the study of the distribution of duration as well as time of occurrence over a given period of time.

3.2 Time of occurrence at multiple time scales

The proposed visual encoding can be customized to show time at multiple scales. The units for the time of occurrence (vertical axis) shown in Fig. 2a can be adjusted to show the time interval $[T_0, T_N]$, as either hours of the day, days of the week, months, or even years.

The use of color adds a further dimension to the visualization. For example the vertical axis can represent the hour of the day, with a range starting from $T_0 = 8am$ to $T_N = 5pm$ at 1-hour intervals; whilst data points can be simultaneously colored according to the day of the week (i.e. data points in red occur on Mondays and blue on Fridays). Combinations of multiple time scales between the vertical axis and the data point color are possible.

3.3 Levels of detail in the duration of events

In terms of duration, events are typically categorized into either point or interval events [17].

The proposed event box visualization can be customized to show, hide or modify the detail of data points. This allows to produce the following levels of detail (see Fig. 2b):

- *Point event*: event details are fully *collapsed* to reduce visual clutter.
- *Interval event without duration outliers*: duration outliers (if any) are hidden.
- *Interval event with duration outliers*: outliers (if any) are shown.
- *Interval event, detailed quartiles*: data points inside colored sub boxes are shown.
- *Interval event, no detailed quartiles*: data points inside colored sub boxes are hidden.
- *Interval event not colored by time scale*: Data points are colored using transparency, so that the volume of the points is observed.

The above will be used to help reducing the visual clutter during the analysis of data that will be discussed in the following section. Transforming events into point or interval events allow to focus the analysis of duration in selected events. Sometimes the user might be interested in the occurrence of an event without being interested in its duration. A point event will provide context without having to remove it from display.

4 Time Attributes Across Sequences

Our methodology is able to seamlessly integrate the extraction and interactive visualization of sequential and time patterns from temporal event data.

4.1 Sequential and Time Patterns overview

Sequential and time patterns are integrated in a single overview, which is built in four steps: 1) the unique sequences that explain most of the variability in the original dataset are selected, resulting in the removal of unique sequences with relatively low frequency; 2) the selected unique sequences are sorted by similarity using complete-link agglomerative hierarchical clustering [9] and the Levenshtein edit distance [12]; 3) sequences are aligned by multiple events as selected by the user; 4) the events in each unique sequence are represented using the visual encoding in Fig. 2a. This type of overview allows comparison of time attributes across sequences. Fig. 1 presents three configurations of this overview for the same dataset using different visualization settings and filters.

4.2 Multiple alignment and user interaction

The present methodology allows users to interact with the data and focus their analysis on common patterns or on anomalous situations that might be of interest. A variety of interaction mechanisms are offered to manipulate and explore the overview of Sequential and Time Patterns:

- *User-driven multiple alignment:* When new alignment events are selected, sequences are re-sorted by the similarity to the alignment events and by similarity within sequences.
- *Manipulation of event encoding to reduce visual clutter:* User can change the level of detail of selected events as per Sect. 3.3.
- *Filter by date:* Records can be filtered by date range or specific days of the week. The data points described in Fig. 2a can be colored according to the day of the week.
- *Breaking down a unique sequence:* A unique sequence can be subdivided according to the day of the week or other criteria.

5 Case Study and Taxonomy of findings

To demonstrate the capabilities of our methodology, we selected a dataset in the healthcare domain as our case study. From the analysis of this dataset, four main classes of findings were identified. We propose a taxonomy of findings which can be generalized to other application domains.

5.1 Case study

The understanding of patient flow is an area where good data analysis is critical [1]. Waiting times, lengths of stay and clinical pathways are key aspects to the study of patient flow.

In this work, we used one year of real-world patient flow data from a Rheumatology outpatient clinic (Sheffield Teaching Hospitals - NHS Foundation Trust, Sheffield, United Kingdom). On average, the Department has an approximate annual load of 9000 patients and 25000 appointments. Patient visits at this clinic are routinely tracked using an in-house workflow tracking system, where the clinical staff (e.g nurses, receptionist, consultants) input the current state of a patient according to the service being provided. The produced event logs are used by the hospital to obtain basic statistics about the quality of care being delivered, particularly focusing on the study of waiting times and lengths of visit.

Our analyses offered the possibility of delving into the raw event logs to extract key insights about patient flow within the clinic (see specific dataset details in Table 1). These are being used to gain a better understanding of how the department operates and to suggest strategies about how to optimize the delivery of care.

Table 1: Characteristics of the Rheumatology dataset.

No. event types	Total no. sequences	No. unique sequences	Time period
18	26,455	161	1 year

5.2 Taxonomy of findings and examples

The generic taxonomy of findings is presented in this section. This taxonomy comprises four classes of findings. Examples for each class are provided using the Rheumatology dataset.

Class I - Discovering main sequential patterns: The user can visually identify the main and most representative sequential patterns in the Sequential and Time Patterns overview. Examples of this class are:

- The most frequent unique sequence is the clinical pathway Appointed → Check-in Kiosk → Height and Weight → Waiting Consultation → In Consultation → Waiting Blood Room → In Blood Room → Complete (Finding 1 in Fig. 1a).

Class II - Impact of an event in the duration of subsequent events: The proposed sorting and alignment methodology allows detecting differences and relations among sequences. The variation in the duration of specific events across sequences could be related to the presence or absence of certain events. For example, the sequences "ABCDE" and "BCDFG", share the subsequence "BCD". This class of finding suggests that if the duration of "BCD" varies between the two sequences, that variation might be associated to the occurrence of the event "A". This finding should require further investigation to explain the situation.

Examples of this class are:

- The time of the event In Consultation is considerably longer when the event Height and Weight does not occur in that sequence (Finding 2 in Fig. 1b). When we interviewed the clinical staff, they indicated that those patients, without Height and Weight, visit the clinic to undergo longer day-case procedures rather than conventional consultations. A day-case procedure will likely include a consultation plus extra activities such as intravenous infusion of drugs, X rays and other imaging tests; hence the long duration of this event.

Class III - Trends in duration with regard to time of occurrence: The proposed visualization allows to identify trends of duration through time. For example, if the duration increases as the time elapses or if high durations occur only in the early hours or a specific day. Examples of this class are:

- In Finding 3 in Fig. 1b, the duration of the late arrival event decreases as the day goes by; meaning that the amount of minutes that patients are late are higher in the morning than towards the end of the day. The highest durations (4th quartile and outliers) are concentrated in the morning (between 9:00 and 12:00). This requires a further investigation of morning appointments, revising the reasons for morning late arrivals.

Class IV - Temporal distribution of time attributes: The proposed visualization helps identifying what is the "normal" duration of an event, as well as identifying the distribution of points through time of occurrence. Examples of this class are:

- Distribution of time of occurrence: On Thursdays, the visits to the clinic feature a higher concentration before noon (Finding 4 in Fig. 1c). However, the visits for which the purpose is exclusively a blood test occur in the afternoon (Finding 5 in Fig. 1c).
- Distribution of duration: On Tuesdays, consultation times are significantly shorter, which suggests that the clinic running on that day might be dealing with less complex pathologies (Finding 6 in Fig. 1b).
- Unusual times of occurrence: On Thursdays, the majority of patients visit the clinic in the mornings. Nevertheless, a reduced number of patients are seen in the afternoons (Finding 7 in Fig. 1c). Investigating that cohort of patients would be of interest.
- Unusual duration: The present method allows for the identification of patients that stay in an event for an unusual amount of time. Patients that have stayed in an event for an unusual time are represented as outlier data points in the proposed visualization, these are visually identified as points outside the colored sub boxes (e.g. see highlighted outliers in Fig. 1, Finding 8).

6 Discussion and Conclusion

This paper proposes a generic methodology to analyze time attributes in event data, at event level and across event sequences. A Sequential and Time Patterns overview visualization is built after extracting the most relevant unique sequences. To simplify the exploration of temporal relations among sequences, unique sequences are sorted by similarity and aligned by multiple events selected by the user.

To facilitate the understanding of how events relate across sequences and within sequences, a novel visual encoding of duration and time of occurrence is embedded in the Sequential and Time patters overview. This encoding enables users to gain valuable insights in to how events distribute with regard to time attributes. To make this possible, users are able to interact with the data at multiple levels of detail and time scales using a single overview.

Based on the presented case study in the healthcare domain, a taxonomy of findings generalizable to other applications has been proposed. The four classes of conclusions presented have already opened the possibility of identifying patterns of interest that should be further investigated to support hospital service improvement.

Depending on the application domain, the methodology could present limitations in terms of scalability. Real-world datasets can contain hundreds or thousands of unique sequences, future work is necessary to create summaries of unique sequences at different levels of detail. Other scalability issues could be present as the ranges of duration $Q_4 - Q_0$ and time of occurrence $T_N - T_0$ become larger, or as the volume of the data points in an event box increase. To solve these potential limitations: time windows, zoom and scaling functions could be used.

References

[1] M. Bardsley, A. Steventon, and G. Fothergill. Untapped potential: Investing in health and care data analytics. *The Health Foundation*, 2019.

[2] V. Barnett and T. Lewis. *Outliers in statistical data.* John Wiley & Sons Ltd., 3nd edition edition, 1994.

[3] R. J. C. Bose and W. van der Aalst. Trace alignment in process mining: opportunities for process diagnostics. In *International Conference on Business Process Management*, pages 227–242. Springer, 2010.

[4] B. C. Cappers and J. J. van Wijk. Exploring multivariate event sequences using rules, aggregations, and selections. *IEEE Transactions on Visualization & Computer Graphics*, (1):1–1, 2018.

[5] Y. Chen, P. Xu, and L. Ren. Sequence synopsis: Optimize visual summary of temporal event data. *IEEE transactions on visualization and computer graphics*, 24(1):45–55, 2018.

[6] M. Frigge, D. C. Hoaglin, and B. Iglewicz. Some implementations of the boxplot. *The American Statistician*, 43(1):50–54, 1989.

[7] D. Gotz and H. Stavropoulos. Decisionflow: Visual analytics for high-dimensional temporal event sequence data. *IEEE transactions on visualization and computer graphics*, 20(12):1783–1792, 2014.

[8] S. Guo, K. Xu, R. Zhao, D. Gotz, H. Zha, and N. Cao. Eventthread: Visual summarization and stage analysis of event sequence data. *IEEE transactions on visualization and computer graphics*, 24(1):56–65, 2018.

[9] J. A. Hartigan. Clustering algorithms. 1975.

[10] B. C. Kwon, J. Verma, and A. Perer. Peekquence: Visual analytics for event sequence data. In *ACM SIGKDD 2016 Workshop on Interactive Data Exploration and Analytics*, volume 1, 2016.

[11] H. Lam, D. Russell, D. Tang, and T. Munzner. Session viewer: Visual exploratory analysis of web session logs. 2007.

[12] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[13] Z. Liu, B. Kerr, M. Dontcheva, J. Grover, M. Hoffman, and A. Wilson. Coreflow: Extracting and visualizing branching patterns from event sequences. In *Computer Graphics Forum*, volume 36, pages 527–538. Wiley Online Library, 2017.

[14] Z. Liu, Y. Wang, M. Dontcheva, M. Hoffman, S. Walker, and A. Wilson. Patterns and sequences: Interactive exploration of clickstreams to understand common visitor paths. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):321–330, 2017.

[15] M. H. Loorak, C. Perin, N. Kamal, M. Hill, and S. Carpendale. Timespan: Using visualization to explore temporal multi-dimensional data of stroke patients. *IEEE transactions on visualization and computer graphics*, 22(1):409–418, 2016.

[16] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. Temporal event sequence simplification. *IEEE transactions on visualization and computer graphics*, 19(12):2227–2236, 2013.

[17] M. Monroe, K. Wongsuphasawat, C. Plaisant, B. Shneiderman, J. Millstein, and S. Gold. Exploring point and interval event patterns: Display methods and interactive visual query. *University of Maryland Technical Report*, 2012.

[18] A. Perer and F. Wang. Frequence: interactive mining and visualization of temporal frequent event sequences. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 153–162. ACM, 2014.

[19] J. W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, 1977.

[20] K. Vrotsou, J. Johansson, and M. Cooper. Activitree: Interactive visual exploration of sequences in event-based data using graph similarity. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):945–952, 2009.

[21] T. D. Wang, C. Plaisant, B. Shneiderman, N. Spring, D. Roseman, G. Marchand, V. Mukherjee, and M. Smith. Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. *IEEE transactions on visualization and computer graphics*, 15(6):1049–1056, 2009.

[22] K. Wongsuphasawat and D. Gotz. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2659–2668, 2012.

[23] J. Zhao, Z. Liu, M. Dontcheva, A. Hertzmann, and A. Wilson. Matrixwave: Visual comparison of event sequence data. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 259–268. ACM, 2015.