Dynamic Hierarchical Aggregation, Selection Bias Tracking, and Detailed Subset Comparison for High-Dimensional Event Sequence Data

Jonathan Zhang* UNC-Chapel Hill
David Borland UNC-Chapel Hill
Wenyuan Wang UNC-Chapel Hill
Joshua Shrestha UNC-Chapel Hill
David Gotz† UNC-Chapel Hill

ABSTRACT

With the increase in collection of temporal event data, especially electronic health record (EHR) data, numerous different visualization and analysis techniques have been developed to assist with the interpretation of such data. As datasets grow increasingly large in both number of event sequences and number of event types, two problems arise: how to group event types, and how to describe selection bias that can occur when selecting cohorts. This poster summarizes two papers, conditionally accepted to VAST, that introduce a dynamic and interactive algorithm for hierarchical event grouping, a scented scatter-plus-focus visualization that supports hierarchical exploration, a tree-based cohort provenance visualization, and a set of visualizations that provide per-dimension selection bias information for pairs of cohorts [2,4]. These methods are integrated into the web-based interactive medical analysis tool *Cadence*

Index Terms: Temporal event sequence visualization, visual analytics, hierarchical aggregation, medical informatics, high-dimensional visualization, cohort selection, selection bias

1 INTRODUCTION

Across a large range of domains, the collection and analysis of large temporal event data has become more common. Visual analytics techniques are used to aggregate and visualize large collections of event sequences in order to enable users to gain insights about the intricacies of the data. Medical data analysis has been a commonly studied application of event sequence visualization techniques [7] in which there are a high number of distinct event types.

Two problems that arise from data of this type are: (1) the large number of events prevents users from effectively interpreting the data, and (2) filtering of events can cause selection bias in which the group being examined is not representative of the whole population [5]. Both issues can be addressed, but current methods have their limitations. The first issue can be addressed through the process of grouping similar event types together. Current methods group event types together as a pre-process [3], which limits a user's ability to view the data at different levels of aggregation and forces assumptions to be made about the aggregation of data. The second issue can be addressed through contextual visualization methods, that allow the user to be able to assess the "drift" of each individual variable [1]. Current contextual visualization methods do not enable bias comparisons required by many fields [6] due to considering only linear sequences of cohort selection steps and not addressing hierarchical relationships.

This poster summarizes two papers that have been conditionally accepted to VAST [2, 4]. The key research contributions that are presented are:

- An algorithm to select the most informative set of event groups in a given context in order to address the high-dimension of event types and a visualization to display these groupings with tools to allow a user to explore alternative event groupings
- A tree-based cohort provenance visualization that shows the selection process and the selection bias at each step, and a visualization that displays the selection bias between pairs of cohorts
- Integration of these methods within *Cadence*, a temporal event sequence web-based visual analytics and cohort selection tool.

We present these methods integrated in the *Cadence* system through use cases with medical data.

2 USE CASE

Figure 1 and Figure 2 are screenshots of the *Cadence* system that were taken during an analysis of a dataset comprising 1732 patients who were discharged from the hospital after being previously diagnosed with any form of pain. This data that was selected included medical information for up to one year prior to the diagnosis for pain. This data set was used to analyze opiate related disorders in patients that have been discharged with a pain diagnoses.

The *Cadence* system has two main panels that assist the user in analysis. Figure 1 displays the timeline panel, which provides the user with summary statistics, a timeline, and an interactive scatter-plus-focus visualization to assist in hierarchical exploration of event groups. Figure 2 displays the cohort panel, which provides a tree-based provenance visualization of how cohorts were created, an icicle plot of variable drift between cohorts, and a dot plot to further assess drift.

2.1 Timeline Panel

In Figure 1, the user decides to analyze the period of time between the pain diagnosis and hospital discharge. Clicking on the corresponding timeline segment updates the scatter-plus-focus plot on the right and the summary statistics on the left. An algorithm determines a level of aggregation for event groupings in the plot, which can be adjusted by the user with the slider above the plot. We can notice that the event group of *Nicotine dependence* is an outlier relative to other groups of similar hierarchical aggregation. This event is also has a relatively high correlation with opiate disorders ($\rho = 0.13$) and was fairly prevalent in the selected cohort population (360 of 1,732 patients). The user is able to click on this circle representing *Nicotine dependence* to switch a focused mode (not shown). In the focused mode, only events that were a supertype of the selected node(up to the root) or a direct child of the selected node will be displayed(with displaying the super type relationships). This allows the user to see how larger or smaller groupings might be more appropriate for use in their analysis. In this case of *Nicotine dependence*, we see that a supertype node that corresponds to a broader form of substance abuse will have higher correlation and might be more applicable to the analysis depending on context.

This focused view also provides other features that assist the user in their analysis. Since, child nodes commonly group together in the visualization, cluttering is an issue. An optimization based layout is employed to space out events more evenly. Furthermore, a scenting feature is utilized to provide the user with a general idea of the heterogeneity of subtypes (with regard to correlation). This is visually represented as a smaller gylph below nodes.

*e-mail: jzhang42@live.unc.edu
†e-mail: gotz@unc.edu

IEEE VAHC 2019
20 October, Vancouver, BC, Canada
978-1-7281-2019-5/19/$31.00 ©2019 IEEE

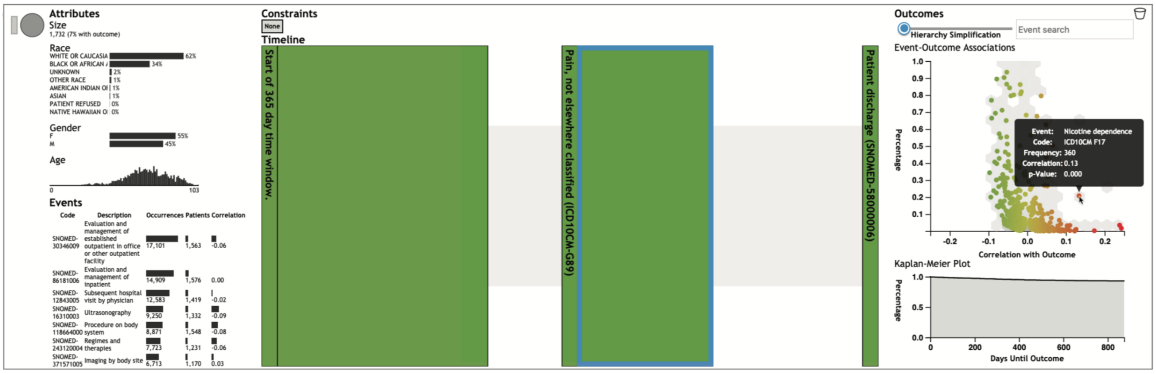

Figure 1: Screenshots from the use case described in section 2.1.

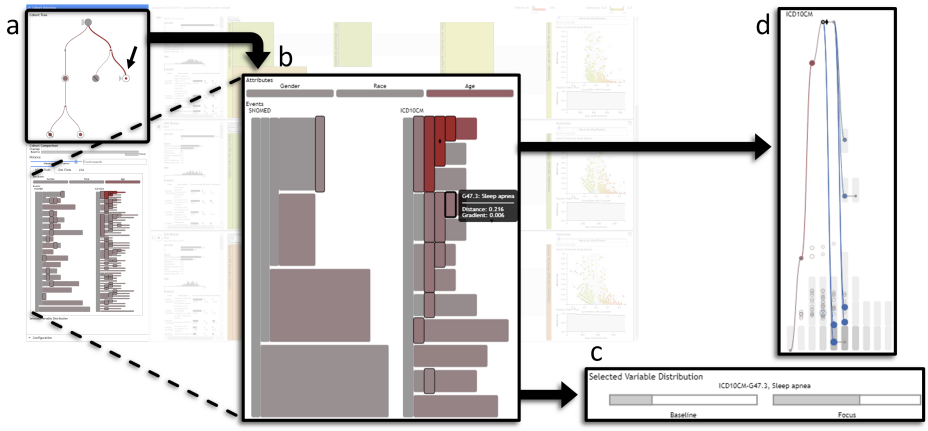

Figure 2: Screenshots from the use case described in section 2.2.

The user is able to click on any node in the focused view to readjust to focus to that node or select any new timeline segment to update the scatter-plus-focus view and summary statistics. The user is also able to add event type groups as new milestones. This results in an updated timeline which contains both the original timeline and new timelines that has the milestones.

2.2 Cohort Panel

In Figure 2a, the user has created numerous cohorts using the interface in Figure 1. The focused cohort contains 227 patients and is filtered by *Obesity and other hyperalimentation*. The size of the inner circle for a glyph is proportional to the cohort size, while the color is indicative of the drift(with gray being less drift and red being more). The red color of this cohort glyph indicates it has drifted further than the other cohorts. To check for differences between the baseline cohort and the selected cohort, the user examines the split icicle plot in Figure 2b which displays the areas of the dimension hierarchy that contribute the most to the high drift. The constrained *Obesity...* variable, indicated by a symbol, has drifted the most.

The user also examines the event with the highest drift in a different branch of the hierarchy and identifies it to be *Sleep apnea*. The selected variable distribution visualization in Figure 2c allows the user to check if patients in the *Obesity...* cohort have a higher prevalence of *Sleep apnea*. The obesity cohort displays a higher rate of *Sleep apnea* than the baseline cohort.

In Figure 2d, the user can examine the drift of descendant variables of the constrained variable identified with a hierarchical dot plot. In this example four descendants maintain a relatively high drift, while the remaining have much lower drift.

3 Conclusion and Future Work

This poster describes the hierarchical aggregation methods and selection bias tracking methods from two papers conditionally accepted to VAST [2,4]. The work done has been integrated into the *Cadence* system and addresses limitations that previous methods have encountered. Future work includes visualization methods that enable for more flexible groupings of events beyond those defined within a specific type hierarchy and providing references to what can be considered as a dangerous amount of drift for a variable.

References

[1] D. Borland, W. Wang, and D. Gotz. Contextual Visualization. *IEEE Computer Graphics and Applications*, 38:17–23, 2018. doi: 10.1109/MCG.2018.2874782

[2] D. Borland, W. Wang, J. Zhang, J. Shrestha, and D. Gotz. Selection bias tracking and detailed subset comparison for high-dimensional data. *Conditionally Accepted to IEEE VAST*, 2019.

[3] F. Du, B. Shneiderman, C. Plaisant, S. Malik, and A. Perer. Coping with Volume and Variety in Temporal Event Sequences: Strategies for Sharpening Analytic Focus. *IEEE Transactions on Visualization and Computer Graphics*, 23(6):1636–1649, June 2017. doi: 10.1109/TVCG.2016.2539960

[4] D. Gotz, J. Zhang, W. Wang, J. Shrestha, and D. Borland. Visual analysis of high-dimensional event sequence data via dynamic hierarchical aggregation. *Conditionally Accepted to IEEE VAST*, 2019.

[5] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins. A structural approach to selection bias. *Epidemiology (Cambridge, Mass.)*, 15(5):615–625, Sept. 2004.

[6] D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340:c869, Mar. 2010. doi: 10.1136/bmj.c869

[7] W. Raghupathi and V. Raghupathi. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2, Feb. 2014. doi: 10.1186/2047-2501-2-3