Visualizing Temporal Patterns by Clustering Patients

Grace Shin, MS[1]; Samuel McLean, MD[2]; June Hu, MS[2]; David Gotz, PhD[1]
[1]School of Information and Library Science; [2]Department of Anesthesiology
University of North Carolina-Chapel Hill, Chapel Hill, NC, USA

Abstract

Medical institutions and researchers frequently collect longitudinal data by conducting a series of surveys over time. Such surveys generally collect a consistent and broad set of data elements from large sets of patients at predefined time points. In contrast to the sparse and irregular retrospective observational data found in electronic medical record (EMR) systems, prospectively gathered survey data captures the same variables at the same time steps across the full study population. Most analyses of this type of longitudinal data focus on understanding the how various properties of the patient cohort associate with specific variables or outcomes measures. However, this approach may miss interesting patterns within constellations of correlated variables. In this paper we describe a visual analysis method for survey data that considers interactions across the full, high-dimensional set of collected variables. Our approach first applies cluster analysis algorithms to survey data collected at each time point independently. We then visualize patient cluster dynamics over time, allowing investigators to identify common patient subgroups and evolution patterns, inspect derived statistical summaries, and compare findings between patient subgroups. We demonstrate our method using data from a survey that followed a cohort of approximately 1,000 patients admitted to the emergency department (ED) following a motor vehicle accident. The survey includes data for each patient at four discrete time points, beginning at admission to the ED and continuing for one year.

1. Introduction

As health information technology becomes more pervasive, institutions are collecting an ever-growing amount of data about the patient experience. In addition to volumes of retrospective electronic health records (EHRs), a significant amount of information is also being gathered via prospective studies designed to collect a specific set of data over time from targeted populations. In contrast to the sparse and irregularly observed data found in EHRs, prospective surveys typically produce dense and consistent sets of data that capture the same data at the same time points for all participants. This provides a rich resource for those seeking to understand temporal, population-level trends in outcomes of interest. Most often, analyses of data from these study focus on understanding of how various properties of the patient cohort associate with specific variables or outcomes measures. While this approach can be highly informative, it may also miss interesting and harder-to-find patterns that are diffused across constellations of correlated variables. In addition, those interested in understanding the data have no ability to explore outcomes and relationships interactively.

This paper describes an interactive visual analysis method designed to help discover and highlight such hard-to-find patterns. Our approach applies user-configurable cluster analysis algorithms to participant data independently at each time step. This produces a set of multiple cohort segmentations, one for each time period. We then visualize changes in patient cluster membership over time, capturing the aggregate dynamics of how participants evolve from time step to time-step including common patient subgroups and transitions. Interaction capabilities allow users to inspect derived statistical summaries for specific cohorts, and compare findings between patient subgroups.

These methods draw on a rich history of work exploring temporal visualization of patient medical data as we describe in the Related Work section. Particularly relevant are flow-based diagrams that show, as our method does, aggregate cohort evolution patterns over time[1–4]. These techniques have shown that graphical visualizations of patient data arranged temporally (in timeline fashion) can provide a useful way for physicians to view the progression of sets of patients. However, these methods typically focus on visualizing low-level medical events such as individual diagnoses or medications. Unfortunately, medical data is of such high dimensionality that the number of variations is very large. Moreover, small variations

in time or sequence that may not be clinically significant can significantly alter the results. For these reasons, more general higher-level trends are often difficult to uncover. Our method, because it focuses on survey data that are from specific time points, takes a different approach. Rather than plotting specific medical events, our system identifies and visualizes clusters of similar but not identical patients. This allows high-level pattern identification and analysis that overcomes challenges of scale that occur due to small variations in underlying patient data.

To validate our approach, we applied our method to patient survey data from a survey of approximately 1,000 patients who were injured in a vehicular accident and required treatment at an emergency room. The survey captured a wide variety of data from these patients at four discrete time steps: arrival at the emergency department (ED), six weeks later (W6), six months later (M6), and one year later (Y1)[5].

We developed an interactive visualization prototype based on our methods and used it to (1) analyze the survey data to identify clusters of similar patients at each of the four time points, (2) visualize patient trajectory between clusters over time, and (3) support interactive exploration and comparison of descriptive statistics calculated for each dynamically computed patient cluster. The prototype supports a range of clustering algorithms and parameter controls, allowing for exploration of different types of patient groupings.

2. Related Work

Given the central role of time in many medical datasets, temporal visualization methods have been used in many different medical informatics contexts. For example, a number of systems have adopted visualization as a means to convey data for individual patients. For example, Plaisant et al. developed LifeLines[6] which provides a timeline-based visualization environment for personal patient medical histories. Similarly, Powsner and Tufte developed a graphical summary of patient status using a table of individual plots of treatment data and test results[7]. As a final example, TimeLine by Bui et al.[8] outlines another variation of vertically arranged timelines representing an individual patient's data.

Recognizing the importance of understanding population-level dynamics, a number of more recent research efforts have proposed visualization methods designed for depicting data for sets of patients. Fails et al. developed PatternFinder[9], an interface that provides result-set visualizations to search for and discover temporal patterns within multivariate datasets which was applied to analyze patients with high blood sugar. Meanwhile, Wang et al.[10] presented an interactive visual tool to visually align sets of individual patient timelines around sentinel events through which patients exhibiting specific event sequences could be found.

While the examples above support the visual analysis of data from multiple patients, they achieve this through small multiples: repeated graphical elements that individually represent each patient. Large-scale cohorts—with hundreds, thousands, or even millions of patients—pose a difficult challenge for this approach. For that reason, scalable flow-based visualization techniques have been used to depict patient evolution in aggregate. Examples of this approach include LifeFlow[3], Outflow[2,11], and DecisionFlow[4]. These techniques all use individual medical events (e.g., a single diagnosis or medication event) to group patients into a single flow. In this paper, we propose an alternative method that displays clusters of statistically similar patients who might not share identical event sequences in their record.

Our approach is certainly not the first to use statistical analysis to group patients and visualize the results. Patient stratification has long been used to prioritize patient populations or identify those most at risk[12]. However, previous methods have differed from those presented here in that they have not applied cluster analysis algorithms independently at different periods of time, have not visualized these changes in cluster membership, and/or have not allowed users to interactively explore the data by selecting characteristics upon which to cluster patients or specific subgroups to cluster.

3. Visual Analysis Methods

Our visual analysis method begins with raw study data as input and produces an interactive visualization of patient cohort evolution over time as output. This section describes the key steps in the process of

converting the input data to this final visual display. As shown in Figure 1, these steps include data normalization, variable selection, cluster algorithm configuration, cluster analysis, and visualization.

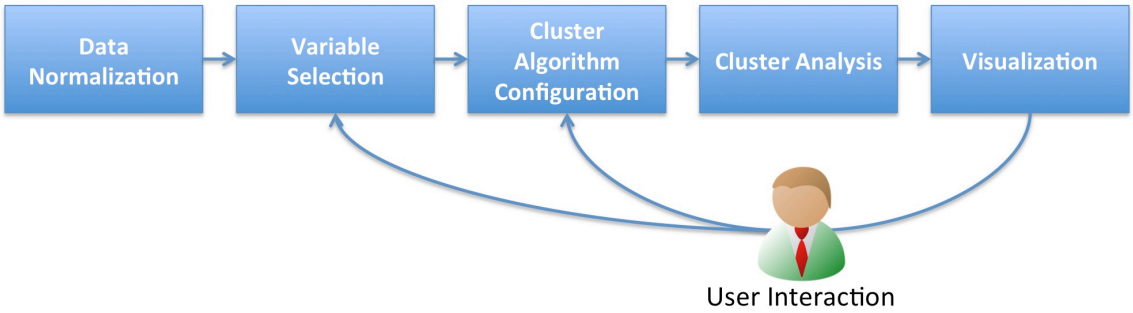

Figure 1. Our method begins with raw survey data which is then normalized and prepared for cluster analysis. Users can optionally select a subset of survey variables (all variables are used by default) and cluster algorithm parameters to use for computing the clusters at each time step. The clusters are then visualized, allowing users to (a) explore changes in cluster membership across time steps and (b) compare summary statistics for each cluster.

Data Normalization. Because the approach outlined in this paper is designed for prospective study data, it assumes that the input data is both dense (that all variables are populated for all patients) and temporally aligned (that measurements are captured for all patients at the same time points after alignment). In practice, however, some data cleaning is often required to omit (or impute) missing values and to clean up other data anomalies. In addition, a data normalization process is required to convert measurements captured using different scales into comparable measurements. This is an important pre-processing step and is necessary to obtain valid results from the cluster analysis algorithms.

Variable Selection. By default, the proposed method clusters patients at each of the datasets time points using all available variables. However, it is often desirable to focus the clustering algorithm on specific subsets of the variable space. For example, an investigator may wish to omit demographic data from his/her analysis. A variable selection panel in the user interface supports this function by allowing users to check (or uncheck) certain variables dynamically over the course of an analysis. The checked variables are considered selected, and only the selected variables will be considered by the system when applying the clustering algorithm. By making this control interactive and part of the user interface, ad hoc exploration patterns are supported. Users can change the variable selection, quickly see the impact of this change on the visualization, and then follow up with additional changes to the selected set of variables. While the user can in theory select any of the available variables (several hundred in the dataset used here), the user interface in our prototype implementation provides a short list of clinically interesting variables selected by content experts. More specifically, we focus on nine variables including four demographic factors and five pain symptom measures.

Cluster Algorithm Configuration. In addition to controlling the set of selected variables used in the cluster analysis, users can configure the clustering algorithm itself. This includes both a selection of the algorithm used and any associated input parameters required by the selected algorithm. Our prototype implementation supports four distinct clustering algorithms. Two methods, Ward's Method for hierarchical clustering[13] and K-Means clustering[14], allow the user to specify the number of clusters to identify at each time point in the study data. The two other supported clustering methods take tuning parameters that indirectly control the degree of clustering as a function of the underlying data distributions: DBSCAN[15] and Affinity Propagation[16]. By providing flexibility in configuring the algorithm and parameters used during clustering, our method allows users to explore the differences in patterns identify by the various algorithms.

Cluster Analysis. The preceding three steps—data normalization, variable selection, and algorithm configuration—prepare the inputs required to perform the actual cluster analysis computations. Based on the specified algorithm configuration and the set of selected variables, the normalized participant data is processed to generate a multiple sets of cluster assignments. One set of clusters is independently computed for the entire patient population at each time point in the data.

Visualization. Once the clusters have been computed for each time step, the results are visualized for interpretation and visual analysis. Our visualization design is patterned after a Sankey diagram and shows both the computer clusters and patients' changes in cluster memberships between time periods. More details of our visual design are described in the next section.

4. Visual Design

The visualization component of our system adopts a flow-based design that builds on traditional Sankey Diagrams. In our design, we represent the individual time points in the study data as a vertical line arranged horizontally across the screen. At each time point, blue rectangles are used to represent individual clusters of patients as computed by the methods described in the previous section. This design is reflected by the vertical blue rectangles in Figures 2 and 3. The height of each blue rectangle corresponds to the fraction of the overall population that belongs to the corresponding cluster. Larger clusters have taller blue rectangles. The cluster rectangles at one time point are connected those at neighboring time points via gray, curving edges. Each edge represents a set of patients that move from one cluster at a particular time to another cluster at the subsequent time step. As with the blue cluster rectangles, the height of each gray curing edge corresponds to the number of patients.

Interaction plays a critical role in the user interface (UI) design. The visualization is placed in a central canvas areas surrounded by two sidebars. The leftmost sidebar contains the variable selection controls and the clustering algorithm configuration controls. These allow user feedback to flow back to earlier stages of the method as illustrated in Figure 1. Once making a set of modifications has been made, users can click on the "Update Visualization" button to trigger a new round of clustering computations based on the current settings in the user interface. As the computation completes, the visualization is updated to reflect the resulting change in patient cluster assignments. This change is reflected in the differences between Figures 2 and 3. Both figures show a visualization of the same underlying patient data and are processed by the same clustering algorithm. Only the lists of selected variables are different between the two screenshots.

In addition, users can mouse over both edges (the grey areas) and nodes (the blue rectangular areas) to learn more about the corresponding participants. Details such as the number of participants and cluster labels are included in the provided data. Finally, users can select an edge to see dynamically computed statistics from the corresponding cluster. The statistical summary, visible in Figures 2 and 3, shows mean values for variety of features and other descriptive statistics. By clicking one by one on the edges in the visualization, users can compare and contrast the profiles of different patient subgroups and begin to learn what participant factors might associate with the progression patterns seen in the Sankey-based visualization.

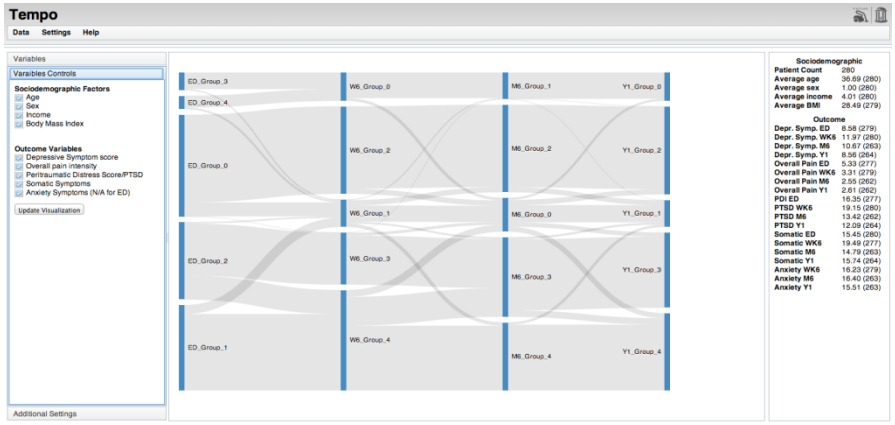

Figure 2. A screen capture of our prototype implementation applied the study data. The four vertical lines of blue rectangles correspond to the four time steps in the data: ED, W6, M6, and Y1. The left sidebar shows the systems variable selection controls while the right sidebar shows detailed statistics for the selected group of participant.

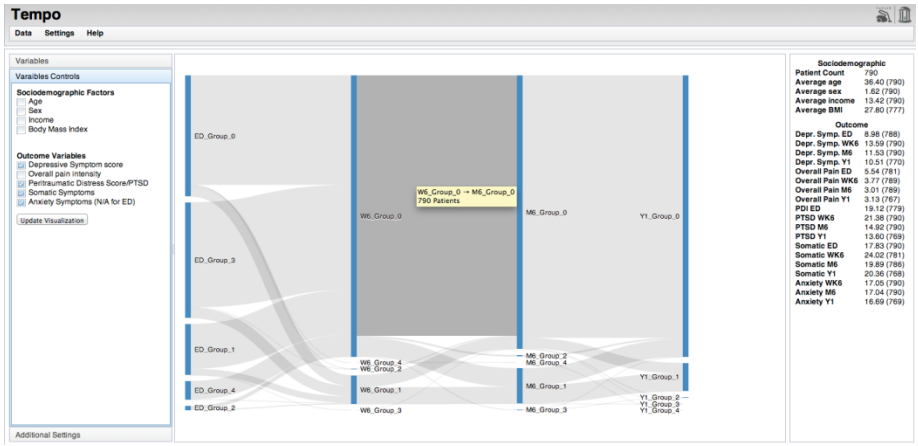

Figure 3. An alternative view of the same data being visualized in Figure 2. The differences in pathways in the visualization are caused directly by the user's interaction with the variable controls panel.

5. Conclusions

This paper described a visual analysis method designed to uncover patterns of participant evolution in longitudinal survey data. Our approach applies cluster analysis algorithms independently to the subsets of survey data collected at each time step. We adopt a Sankey-based visualization design to illustrate participant cluster dynamics over time. Interactions are supported, allowing investigators to identify common participant subgroups and evolution patterns, inspect derived statistical summaries, and compare findings between participant subgroups. We demonstrated our method using data from a 1-year survey capturing data about pain for roughly 1,000 participants who were admitted to the emergency department (ED) following a vehicular accident. We demonstrate how our methods can be applied to this dataset and show examples highlighting the types of analyses that our approach supports.

References

1. Perer, A. & Wang, F. Frequence: Interactive Mining and Visualization of Temporal Frequent Event Sequences. *Proceedings of the 19th International Conference on Intelligent User Interfaces* 153–162 (ACM, 2014).
2. Wongsuphasawat, K. & Gotz, D. Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 2659–2668 (2012).
3. Wongsuphasawat, K. *et al*. LifeFlow: Visualizing an Overview of Event Sequences. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 1747–1756 (ACM, 2011). doi:10.1145/1978942.1979196
4. Gotz, D. & Stavropoulos, H. DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics* Early Access Online, (2014).
5. Platts-Mills, T. F. *et al*. Using emergency department-based inception cohorts to determine genetic characteristics associated with long term patient outcomes after motor vehicle collision: Methodology of the CRASH study. *BMC Emergency Medicine* 11, 14 (2011).
6. Plaisant, C. *et al*. Visualizing Medical Records with LifeLines. in *CHI '98 Conference Summary on Human Factors in Computing Systems* 28–29 (ACM, 1998). doi:10.1145/286498.286513
7. Powsner, S. M. & Tufte, E. R. Graphical summary of patient status. *Lancet* 344, 386–389 (1994).
8. Bui, A., Aberle, D. R. & Kangarloo, H. TimeLine: Visualizing Integrated Patient Records. *IEEE Transactions on Information Technology in Biomedicine* 11, 462–473 (2007).

9. Fails, J. A., Karlson, A., Shahamat, L. & Shneiderman, B. A Visual Interface for Multivariate Temporal Data: Finding Patterns of Events across Multiple Histories. in *Visual Analytics Science And Technology, 2006 IEEE Symposium On* 167–174 (2006).
10. Wang, T. D. *et al*. Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 457–466 (ACM, 2008).
11. Wongsuphasawat, K. & Gotz, D. Outflow: Visualizing Patient Flow by Symptoms and Outcome. in *IEEE VisWeek Workshop on Visual Analytics in Healthcare* (2011).
12. Gotz, D., Stavropoulos, H., Sun, J. & Wang, F. ICDA: A Platform for Intelligent Care Delivery Analytics. in *AMIA Annual Symposium Proceedings* (2012).
13. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58, 236–244 (1963).
14. MacQueen, J. Some methods for classification and analysis of multivariate observations. in (The Regents of the University of California, 1967).
15. Ester, M., Kriegel, H., S, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of KDD* 226–231 (AAAI Press, 1996).
16. Frey, B. J. & Dueck, D. Clustering by Passing Messages Between Data Points. *Science* 315, 972–976 (2007).