Multivariate Visualization of Longitudinal Clinical Data

David Borland, Vivian L. West, and W. Ed Hammond

Fig. 1. Multivariate visualization of longitudinal clinical data related to diabetes, with a selected group of patients highlighted in blue. 1) Scatter plot of features extracted from hemoglobin A1c (HbA1c) values over time for each patient, here showing on the y-axis the slope of a linear regression line fit to the values, and on the x-axis the value mean. 2) Parallel coordinates visualization of multiple temporal features. 3) Hexagonal bin visualization showing the distribution of HbA1c values over time. 4) Icicle plot visualization of ICD-9 diagnosis codes. 5) Parallel sets visualization of demographic data.

Abstract— Identifying different patterns in longitudinal data (e.g. in laboratory values over time) related to a certain disease and associating them with different outcomes can be an important factor in delivering improved patient-specific health care. Visualization is often used to help understand complex temporal patterns, however the effectiveness of many temporal visualization techniques is compromised when dealing with temporal data from large numbers of individuals. We present work in progress on a visualization tool for exploring trajectories in longitudinal clinical data and their relationships to other disease factors. Linked views of multivariate features calculated from the longitudinal data, dynamically aggregated longitudinal data, diagnoses, and demographic data are presented, enabling the user to explore the data and identify potential relationships between temporal patterns, diagnoses, and demographics. We describe the various feature of this tool, demonstrate its application to patients diagnosed with diabetes, and discuss future work.

Index Terms—Temporal visualization, multivariate visualization, longitudinal data, human computer interaction, electronic health records.

1 Introduction

The increasing adoption of electronic health records (EHRs) provides ever-growing sets of rich data that can be analyzed to provide new clinical knowledge and enable more effective health care [7, 17, 19]. Specifically, longitudinal clinical data, such as multiple laboratory values taken at discrete points over time, related to a given disease can help define different temporal trajectories corresponding to different disease progressions. Our hypothesis is that identifying these disease trajectories and their relationships to other factors, such as comorbidities or demographic data, can help define different subpopulations of a disease that may benefit from different treatments.

Visualization is a tool often used to help analyze longitudinal data. However, the effectiveness of many temporal visualization techniques is compromised when dealing with data from large cohorts. For example, simple line graphs of data for thousands, or tens of thousands, of patients can result in overplotting, impeding the interpretation of, and identification and selection of individuals with, temporal characteristics of interest.

- *David Borland is with RENCI, The University of North Carolina at Chapel Hill. E-mail: borland@renci.org.*
- *Vivian L. West is with the Duke Center for Health Informatics, Duke University. E-mail: vivian.west@duke.edu.*
- *W. Ed Hammond is with the Duke Center for Health Informatics, Duke University. E-mail: william.hammond@duke.edu.*

Proceedings of the IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis. Available online at: http://eventevent.github.io

We are therefore developing a visualization tool to enable exploratory visual analysis of longitudinal clinical data for large cohorts of patients that enables the user to identify and select temporal patterns of interest, and see the relationship of these patterns to other clinical data (Fig. 1). We demonstrate the application of this tool to a cohort of 1456 patients diagnosed with diabetes. The primary features of the tool are:

- Multivariate visualizations of features extracted from longitudinal clinical data, enabling the selection of patient groups based on features related to the numerical values of selected lab tests and their distribution in time.
- A visualization of the temporal distribution of the longitudinal data using a 2D binning approach, with optional links between bins to enable direct visualization of temporal patterns.
- Visualizations of diagnoses and demographic data, showing the overall population distribution and concentration of selected patients within each visualization element.
- Linked selection and visualization across all views of the data.

Our tool contains five views of the data, with linked selection across all five views. A scatter plot (Fig. 1-1) and parallel-coordinates plot (Fig. 1-2) provide multivariate visualizations of multiple descriptive features extracted from the longitudinal data, enabling the visualization of relationships between features, and the selection of individuals with certain characteristics. Descriptive features of values (e.g. mean value) and intervals (e.g. maximum period between values) are available. A hexagonal binning visualization (Fig. 1-3) provides an aggregated visualization of the temporal distribution of data points. An icicle plot varient (Fig. 1-4) and a parallel-sets visualization (Fig. 1-5) provide visualizations of diagnoses and demographic data respectively.

In the remainder of the paper we discuss relevant previous work in visualizing longitudinal data (Sect. 2), provide a brief overview of our current dataset (Sect. 3), describe the visualization tool and its constituent views in more detail (Sect. 4), and discuss future work (Sect. 5).

2 Previous Work

Javed et al. [12] compared the efficacy of various visualization techniques for multiple time series applied to general visualization tasks, including standard line graphs, horizon graphs [6, 8, 16], and small multiples [18], although they considered only relatively small numbers of time series. Previous work has identified some of the challenges in the visual analysis of time-oriented data in the healthcare domain, including issues related to scale, complexity, and interaction [1]. Various temporal visualizations of health-related data effectively display sequences of discrete events (e.g. [15,20]), however they do not directly address visualizing continuously-valued time-series data at variable intervals, which is our focus. Related previous work has explored the visualization of longitudinal data related to diabetes [2, 9], both of which employee regular sampling and categorization to enable aggregation and display of large amounts of data, at the expense of losing some information. We hypothesize that temporal patterns of the intervals between lab values may be useful in categorizing different patient groups, so have developed multivariate visualizations to enable the user to select groups based in part on this information.

3 Data Description

The longitudinal data are hemoglobin A1c (HbA1c) values from 1456 patients diagnosed with diabetes, with a total of 27,187 values, or roughly 18 values per patient. HbA1c is a clinical indicator of diabetes control, with higher values related to poorer control. We therefore use HbA1c as a proxy for the temporal trajectory of the disease. Due to variability in the timing of disease diagnosis with respect to disease onset and progression, we have restricted our patient population to those that have died, and aligned the temporal data relative to date of death. Data for up to 16 years prior to death is present. International Statistical Classification of Diseases and Related Health Problems (ICD)-9 codes and demographic data (age at death, gender, and race) are also included.

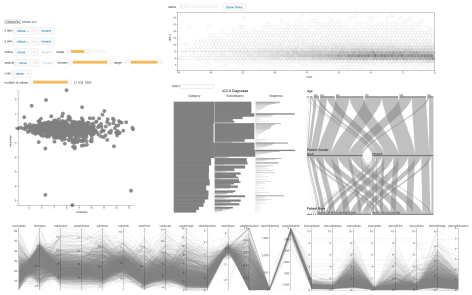

Fig. 2. Visualization with no selection, showing the full patient population in gray.

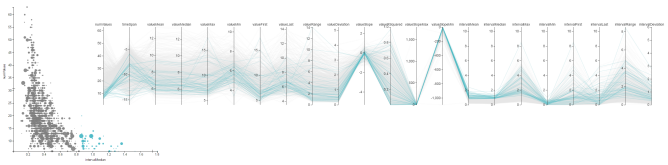

Fig. 3. Scatter plot (left) and parallel coordinates (right), highlighting all patients with a relatively high median interval between HbA1c readings.

4 Visualizations

Our visualization tool is written using D3 [4]. Interactive selection is linked across all views, with selected data represented in blue. Fig. 2 shows the visualization prior to any selection, showing the overall population. A global filter on the number of HbA1c values present per patient can be used to restrict the visualization to only show patients with a certain amount of data. Information on the different views is provided in the following sections.

4.1 Multivariate Temporal Features

To facilitate the selection of patients whose disease trajectories share certain characteristics, we calculate a suite of features (statistical descriptors, linear regression values, etc.) on the HbA1c values themselves, and the intervals between data points. We use a multivariate scatter plot and a parallel coordinates [11] visualization[1] to visualize and interact with these features (Fig. 3), noting that prior work has shown the effectiveness of combining these two types of visualizations in different ways [3, 10, 21]. The user can select which features are mapped to the two axes of the scatter plot, and optionally map other features to point size and opacity. A rectangular selection box can be used to select points of interest. Selection in the parallel coordinates plot is performed by brushing on each axis independently. In practice, the parallel coordinates plot provides a good overview of how the current selection is distributed across all of the calculated features, while the scatter plot provides a more focused view of features of interest.

4.2 Longitudinal Data

Due to the large amount of data being visualized we employ a hexagonal binning approach [5], providing a 2D histogram that directly visualizes the distribution of HbA1c values over time and improves legibility and performance over individual point- or line-based techniques (Fig. 4). The population density of lab values in each hexagonal bin is mapped to grayscale, with darker indicating more values present. The relative density for the current selection is represented by blue hexagons, with density mapped to size. Dashed horizontal lines represent important clinical classifications: normal $< 5.7 \leq$ borderline $< 6.5 \leq$ controlled $< 8.0 \leq$ uncontrolled. The user can select individual bins or groups of bins, and control the bin size for aggregating data. Links connecting bins can also be shown, which can be useful for comparing small groups (Fig. 5).

[1] From: `https://github.com/syntagmatic/parallel-coordinates`

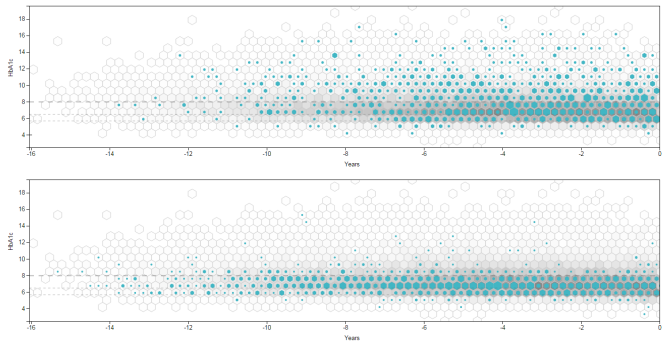

Fig. 4. Hexagonal binning visualization of the temporal distribution of HbA1c values over time. Selected in blue is the distribution of HbA1c values for all patients in their 40s (top) vs. their 90s (bottom) at death, revealing a much greater concentration of high HbA1c values over time for those in their 40s.

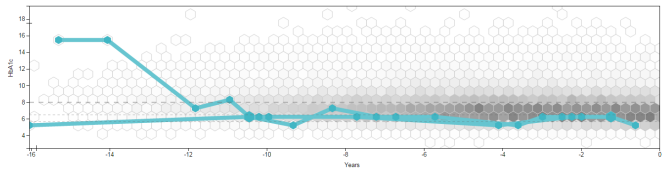

Fig. 5. Links showing trajectories for individual patients.

4.3 Diagnosis Data

ICD-9 codes in our data are organized into a hierarchy with category, subcategory, and diagnosis. We display them with an icicle plot [14] variant which maps diagnosis prevalence to length, width, and intensity (Fig. 6). Initially the overall population prevalences are displayed. Any icicle section can be clicked to select all patients with that diagnosis. Upon selection of a group of patients from any of the views, the same icicle plot structure is maintained to enable ease of comparison, however length is mapped to prevalence within the selection. An autocomplete text-entry field enables quick searching for specific diagnoses, subcategories, or categories. Mouse-over of any icicle when selected enables direct comparison with the overall population prevalence via rectangles with black outlines.

4.4 Demographic Data

To display the available demographic data–age at death, gender, and race–we use a parallel sets [13] visualization[2]. Initially each axis shows the distribution of categories for each demographic dimension, and ribbons show the different combinations of categories across dimensions. Dimensions and categories can be reordered interactively, and categories and ribbons can be selected to highlight all patients in the selected group. Upon selection of a group of patients from any of the views, the same parallel sets structure is maintained to enable ease of comparison, however blue bars for each category, and a light gray to blue color map for each ribbon, are used to convey the prevalence of the respective category or category combination within the selected group (Fig. 7).

5 Discussion and Future Work

Although still a work in progress, use of our tool to explore longitudinal clinical data in diabetic patients has produced some promising initial results. For example, Fig. 8 shows a comparison of a group of patients (group a) whose first and last HbA1c values were high (> 10 in both cases) to a group of patients (group b) whose first HbA1c values had a comparable distribution, but whose last values before death were considerably lower (< 6), selected using the scatter plot view (Fig. 8-1). The longitudinal distribution visualization shows a general trend toward improved control for group b (Fig. 8-2), however comparing the icicle plots shows a noticeable increase in the prevalence of many diagnoses for group b compared to group a (Fig. 8-3). This observation is somewhat counter-intuitive, as improved diabetes control is generally thought to lead to fewer complications. The parallel sets view shows some similarities, primarily with respect to race, and some differences, primarily with respect to gender and age, between the two groups (Fig. 8-4).

[2]Adapted from `https://github.com/jasondavies/d3-parsets`

Fig. 6. Comparison of the prevalence of different diagnoses for patients with a long median interval between HbA1c readings (top) vs. a short median interval (bottom), selected using the scatter plot view (left). The diagnosis prevalences shown in the icicle plot view (right) are noticeably higher for a number of diagnoses in the bottom image, such as those related to certain mental disorders (highlighted in black).

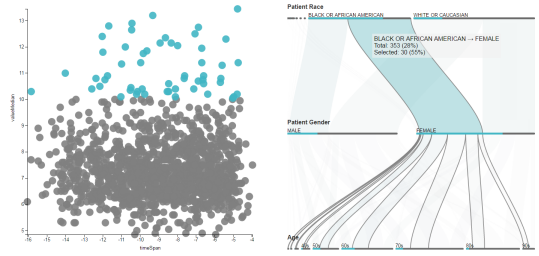

Fig. 7. Selecting all patients with a median HbA1c value ≥ 10 in the scatter plot (left) reveals a high proportion (55%) of black or African American females in this group, as shown in the parallel sets visualization (right).

We have identified a number of avenues for future work. Improving selection via more advanced brushing techniques and the ability to apply set operations to the current selection will enable more fine-tuned exploration of the data. We are also investigating techniques for the selection and visualization of multiple groups to enable improved comparison between groups. Temporal information on dates of diagnoses is not currently integrated into our tool. We will explore different temporal event-based visualizations to incorporate these data and combine them with the temporal data from the HbA1c values. Another avenue for future work will involve investigating different ways of aligning the temporal data such that data from living patients can be included, along with other potential sources such as genetic or environmental data. Incorporating more temporal featurs, and robust statistical methods for determining the significance of perceived relationships in the data, will also be explored.

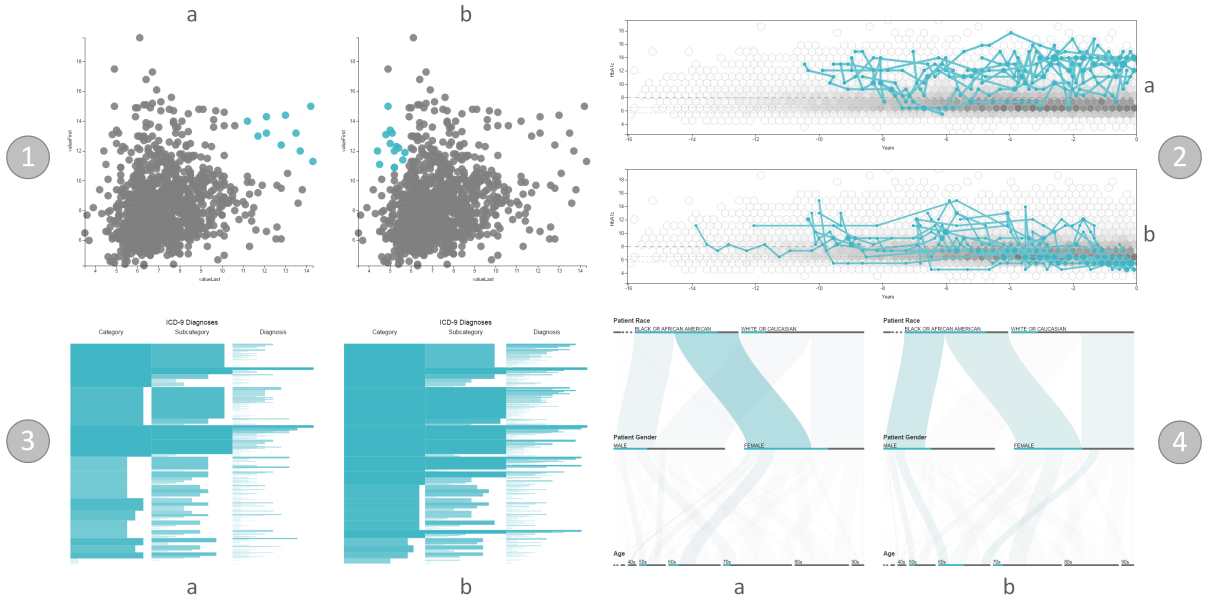

Fig. 8. Comparison of two groups, a and b, of selected patients. 1) The scatter plot is used to select patients with relatively high first (y-axis) and last (x-axis) HbA1c values (a) and patients with comparable first values, but low last values (b). 2) The longitudinal visualization shows a general trend toward improved control for group b vs group a. 3) The icicle plot visualization shows a noticeable increase in the prevalence of many diagnoses for group b vs. group a. 4) The parallel sets visualization shows demographic similarities and differences between the groups.

Acknowledgments

This work was supported in part by the US Army Medical Research and Materiel Command (USAMRMC) under Grant No. W81XWH-13-1-0061. The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other documentation.

References

[1] W. Aigner, P. Federico, T. Gschwandtner, S. Miksch, and A. Rind. Challenges of time-oriented data in visual analytics for healthcare. In *IEEE VisWeek Workshop on Visual Analytics in Healthcare (VAHC)*, Oct. 2012.

[2] D. Borland, E. M. Hinz, L. A. Herhold, V. L. West, and W. E. Hammond. Path maps: Visualization of trajectories in large-scale temporal data. In *Poster Abstracts of IEEE VIS 2015.*, 2015.

[3] D. Borland, W. Vivian L., and W. E. Hammond. Multivariate visualization of system-wide National Health Service data using radial coordinates. In *Proceedings of the 2014 Workshop on Visual Analytics in Healthcare (VAHC 2014)*, 2014.

[4] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, Dec. 2011. doi: 10.1109/TVCG.2011.185

[5] D. B. Carr, R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. Scatterplot matrix techniques for large N. *Journal of the American Statistical Association*, 82(398):424–436, 1987. doi: 10.2307/2289444

[6] S. Few. Time on the horizon. *Visual Business Intelligence Newsletter*, July 2008.

[7] D. Gotz and D. Borland. Data-driven healthcare: Challenges and opportunities for interactive visualization. *IEEE Computer Graphics and Applications*, 36(3):90–96, May 2016. doi: 10.1109/MCG.2016.59

[8] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1303–1312. ACM, New York, NY, USA, 2009. doi: 10.1145/1518701.1518897

[9] E. M. Hinz, D. Borland, H. Shah, V. L. West, and W. E. Hammond. Temporal visualization of diabetes mellitus via hemoglobin A1c levels. In *Proceedings of the 2014 Workshop on Visual Analytics in Healthcare (VAHC 2014)*, 2014.

[10] D. Holten and J. J. van Wijk. Evaluation of cluster identification performance for different PCP variants. In *Proceedings of the 12th Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis'10, pp. 793–802. The Eurographs Association & John Wiley & Sons, Ltd., Chichester, UK, 2010. doi: 10.1111/j.1467-8659.2009.01666.x

[11] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, Aug. 1985. doi: 10.1007/BF01898350

[12] W. Javed, B. McDonnel, and N. Elmqvist. Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):927–934, Nov. 2010. doi: 10.1109/TVCG.2010.162

[13] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568, July 2006. doi: 10.1109/TVCG.2006.76

[14] J. B. Kruskal and J. M. Landwehr. Icicle plots: Better displays for hierarchical clustering. *The American Statistician*, 37(2):162–168, 1983. doi: 10.2307/2685881

[15] 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.

[16] H. Reijner. The development of the horizon graph. In *Electronic Proc. Vis08 Workshop From Theory to Practice: Design, Vision and Visualization*, 2008.

[17] A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, and B. Shneiderman. Interactive information visualization to explore and query electronic health records. *Foundations and Trends in HumanComputer Interaction*, 5(3):207–298, Feb. 2013. doi: 10.1561/1100000039

[18] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, second ed., 2001.

[19] V. L. West, D. Borland, and W. E. Hammond. Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 22(2):330–339, Mar. 2015. doi: 10.1136/amiajnl-2014-002955

[20] 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, Dec. 2012.

[21] X. Yuan, P. Guo, H. Xiao, H. Zhou, and H. Qu. Scattering points in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1001–1008, Nov. 2009. doi: 10.1109/TVCG.2009.179