Proceedings of the 2014

Workshop on Visual Analytics in Healthcare:

November 15th, 2014
Washington, DC
www.visualanalyticshealthcare.org

Sponsors:

IBM Research

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

AMIA Public Health Informatics Working Group

Preface:

As medical organizations move to electronic medical records and embrace new health information technology (HIT), the amount of data available to clinicians continues to grow at a rate not seen before. The vast amount of observational data we now collect promises many benefits. However, there are also many challenges. The large amount of clinical data captured just for a single patient poses a challenging task for clinicians trying to make sense of the patient's condition and understand the patient's medical history. Similar challenges of scale and complexity are faced by those performing population studies over large collections of historical electronic health data.

Visualization and visual analytics show great potential as methods to analyze, filter, and illustrate this vast sea of electronic medical data. By tapping into the human ability to visually perceive and interact with data, these technologies promise to help unlock the hidden insights buried within the large and heterogeneous data sources we are collecting. These insights can help with diagnosis and decision support for physicians. They can help patients and caregivers manage their health. They can help generate and confirm hypotheses for researchers, and they can help institutions improve efficiency, performance, and outcomes.

Visualization and visual analytics can potentially provide great benefits to each of these constituencies. However, to be successful, visualizations must be carefully designed to meet the challenges of the healthcare domain. This event, the fifth annual VAHC Workshop, provides clinicians, medical informaticists, and visualization researchers with a venue to gather, share ideas and results, and discuss directions for future work.

Jesus J Caban, PhD
NICoE, Walter Reed Bethesda
Email: jesus.j.caban.civ at mail.mil

David Gotz, PhD
University of North Carolina
gotz@unc.edu

Adam Perer, PhD
IBM T. J. Watson Research Center
Email: adam.perer at us.ibm.com

Hadi Kharrazi, MD, PhD
Johns Hopkins Bloomberg School of Public Health

Invited Speakers

Kathy Rowell,
Co-Founder of Katherine S. Rowell & Associates and HealthDataViz

Kathy Rowell is co-founder and principal of Katherine S. Rowell & Associates and HealthDataViz, a Boston firm that specializes in helping healthcare organizations organize, design, and present visual displays of data to inform their decisions and stimulate effective action. She advises providers, payers, policymakers and regulatory agencies how to align systems, design reports, and develop staff to communicate healthcare data clearly.

Greg Nelson,
Founder of ThotWave

Greg Nelson is a healthcare executive founder of ThotWave, a consulting services organization that helps organizations get more value out of their data. He has provided direction to projects covering Business Change Management, Process Improvement, Business Intelligence, Analytics, Data Warehousing, Master Data Management, Data Governance, Data Quality, and Research Informatics. He is the author of multiple white papers about Clinical Dashboards including "The Healthcare Performance Dashboard: Linking Strategy to Metrics" and "Building Your First Dashboard Using the SAS 9 Business Intelligence Platform: A Tutorial".

Yair Rajwan,
Founder of Visual Science Informatics

Dr. Yair Rajwan is the founder and director of Visual Science Informatics, a Virginia firm that helps organizations analyze and visualize data and text to provide insight and improve engagement. He is the founder of VisualMatics - Visual Science Informatics, a social network group that connects healthcare practitioners, visual analytics professionals, and information visualization researchers to collaborate on visual analytics proposals and grants.

Agenda:

Session I: Temporal Visualization of Health Data Chair: TBD	
8:30 - 8:45	Welcome – VAHC 2014
8:45 - 10:00	**Paper presentations**
	"Interactive Analysis of Multiple Longitudinal Records of Diabetes Patients" Denis Klimov, Alexander Shknevsky Robert Moskovitch, and Yuval Shahar
	"Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels" Eugenia McPeek Hinz,David Borland,Hina Shah,Vivian West, and Ed Hammond
	" Detecting Novel Associations for Surgical Hospital Readmissions in Large Datasets by Interactive Visual Analytics" Stein Olav Skrovseth, Adam Perer, Conor P. Delaney, Arthur Revhaug, Rolv-Ole Lindsetmo and Knut Magne Augestad
	"Visual Hypothesis & Correlation Discovery for Precision Medicine" Wolfgang Rumpf, Jenn Gonya and William Ray
	JAMIA Publication *"A Novel Tool to Visualize Chronic Kidney Disease Associated Polymorbidity: A 13-year Cohort Study in Taiwan"* Chih-Wei Huang1, Shabbir Syed-Abdul, Wen-Shan Jian, Usman Iqbal, Phung Anh Nguyen, Peisan Lee, Shen-Hsien Lin, Wen-Ding Hsu, Mai-Szu Wu, Chun-Fu Wang, Kwan-Liu Ma, Yu-Chuan (Jack) Li
10:00 - 10:30	Coffee Break

Session II: Clinical Dashboards Chair: TBD	
10:30 - 10:50	*"Transforming Healthcare Data Into Compelling Stories and People Into Compelling Storytellers"* Kathy Rowell, Katherine S. Rowell & Associates and HealthDataViz
10:50 - 11:10	*"Healthcare Analytics: The Next Chapter in Innovation"* Greg Nelson, ThotWave Technologies
11:10 – 11:30	*"Visual Insight for Better Decision: Revealing Meaningful Values of Visual Analytics in Healthcare Dashboards"* Yair Rajwan, Visual Science Informatics, LLC
11:30 – 12:00	**PANEL:** **Introduction to Clinical Dashboards** • Kathy Rowell, HealthDataViz • Greg Nelson, ThotWave Technologies • Yair Rajwan, Visual Science Informatics
12:00 – 1:00	Lunch break

Session III: Poster and Demonstrations Chair: TBD	
1:00 – 1:30	Fast forward Presentations – Posters and Live Demonstrations
1:30 – 3:00	Poster and Demo Session (see next four pages for list of posters & live demonstrations)

Session IV: Visualization of Patterns and Complex Health Data Chair: TBD	
3:00 – 4:30	**Paper Presentations**
	"WorkflowExplorer: Visual Exploration and Identification of Common Multitasking Patterns in Emergency Department Workflow" Allan Fong, Kevin Maloy and Raj Ratwani
	"Visualizing Temporal Patterns by Clustering Patients" Grace Shin, Samuel McLean, June Hu and David Gotz
	"Visualizations of Inter-Observer Reliability Assessments in Time Motion Studies: Facilitating Observers' Training" Marcelo Lopetegui, Po-Yin Yen, Alejandro Mauro, Barbara Lara and Philiip Payne
	"Multivariate Visualization of System-Wide National Health Service Data Using Radial Coordinates" David Borland, Vivian West, and Ed Hammond
	JAMIA Publication *" Seeing the Forest Through the Trees: Uncovering Phenomic Complexity Through Interactive Network Visualization""* Jeremy L Warner, Joshua C. Denny, David Kreda, and Gil Alterovitz
4:30 - 4:35	Closing Remarks

Poster Presentations (1:00 – 3:00pm)	
Poster #1	*"Visual Analysis of Infection Surveillance"* Penny Cooper, Nora Haney and Jose Morey
Poster #2	*"Iterative Methodology and Usability Improvement for Clinical Decision Support"* Fei Yu, Ketan Mane, Vincent Carrasco and Javed Mostafa
Poster #3	*"A Comparative Analysis of Clinical Data Visualization Development Strategies: Build vs. Buy",* Jaehoon Lee, Thomas Oniki, John Holmen, Peter Haug and Stanley Huff
Poster #4	*"Text Mining and Visualization to Explore E-Cigarette and Hookah-Related Social Media",* Annie Chen, Shu-Hong Zhu and Mike Conway
Poster #5	*"Interactive Demographic Visualization of Multiple Facilities across Time"* Hyunggu Jung, E. Sally Lee, Hossein Estiri and Kari A. Stephens
Poster #6	*"Envisaging Biomedicine: Visualization of the Clinical and Research Ecosystem using Bibliographic Metadata"* Terrie Wheeler, Karen Gutzman, Michael Bales, Paul Albert and Kristi Holmes

Live Demonstrations (1:00 – 3:00pm)	
Demo #1	*"Visualization and prediction of diabetes disease progression along with modifiable factors using Diabetes Complications Severity Index (DCSI)"* Muhammad Dastagir, Keith Glassford, Shalom Halevy, Erik Smith and Thomas Van Gilder
Demo #2	*"Weave: Utilization of InfoMaps and an Individual Record Tool for Patient Data Analysis"* John Fallon and Georges Grinstein
Demo #3	*"Demonstration of Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels"* Hina Shah, David Borland, Eugenia Mcpeek Hinz, Vivian L. West and W. Ed Hammond
Demo #4	*"Visual Hypothesis & Correlation Discovery for Precision Medicine"* Wolfgang Rumpf, Jenn Gonya and William Ray
Demo #5	*"EventFlow - Interactive Event Sequence Visualization and Querying"* Fan Du, Sana Malik, Catherine Plaisant, Ben Shneiderman
Demo #6	*"Visual Insight for Better Decision: Revealing Meaningful Values of Visual Analytics in Healthcare Dashboards"* Yair G. Rajwan
Demo #7	*"Understanding Outcome Measures of Patients Diagnosed with mild TBI - A Visual Analytics Approach"* Ryan Diehl, Niki Noprapa, and Jesus J Caban
Demo #8	*"HCC Risk Browser: Visualizing Opportunity & Intervention"* *Michael Simon and Nick Stepro*
Demo #9	*"An interactive Visualization System with a Grammar Induction Layer for Learning and Generating Suggestions from Complex Clinical Datasets"* Filip Dabek, Jesus J Caban, Tim Oates

Proceedings of the 2014
Workshop on Visual Analytics in Healthcare

PAPER PRESENTATIONS

Interactive Analysis of Multiple Longitudinal Records of Diabetes Patients

[1]Denis Klimov, MSc, [1]Alexander Shknevsky, [2]Robert Moskovitch, PhD, [1]Yuval Shahar, MD, PhD,
[1]Medical Informatics Research Center, Department of Information Systems Engineering,
Ben Gurion University, Beer Sheva, Israel
[2]Department of Biomedical Informatics, Columbia University, NYC, USA

Abstract

*To support an interactive process of data analysis and discovery of new knowledge from large volumes of time-oriented clinical data, we had designed and implemented the **Vi**sual **T**emporal **A**nalysis **Lab**oratory (**ViTA-Lab**) framework. The ViTA-Lab framework combines computational data-driven temporal data mining techniques, with interactive, query-driven visual analytical capabilities for investigation of the time-oriented data and of the discovered concepts and patterns. To demonstrate and assess our framework, we explored the data of 1700 diabetic patients, followed episodically over five years, focusing on the data associated with several medications for reducing the level of blood glucose and of its long-term measure, HbA1c (glycated hemoglobin). The exploration clearly shows that Metformin was associated with a significantly higher rate of decreased HbA1c (as well as with a shorter interval needed for achieving that decrease) than Glibenclamide.*

1. Introduction: Application of user-driven and data-driven analysis approaches in the medical domain

The effective interpretation and analysis of time-oriented multivariate data, and in particular, of longitudinal clinical data, embodies within it a significant potential for the discovery of clinically significant medical knowledge, leading to an improvement in the quality of clinical care.

Interactive visual exploration systems provide users with an overview of the data, enabling them to explore the visualized data to answer user-initiated queries. Thus, we refer to this data analysis manner as a *query-driven* (or *user-initiated*) approach. Although the interactive query-driven visual approach is user friendly and is highly focused on the analysis of concepts that have a relatively high promise for the discovery of new, meaningful knowledge, the method has a major inherent limitation: the user must know exactly what to look for, and which questions to ask. For example, if a query about a key association among three concepts has not been asked, a potentially important pattern might never be discovered. Thus, query-driven methods are quite *precise*, in the sense of producing mostly significant answers, but are *incomplete*.

Previously, visual exploration systems in medical domains focused mostly on the visualization of raw longitudinal data for individual [1, 2] or multiple patient records [3, 4], as reviewed by Chittaro [5] and recently by Rind et al [6]. The seminal review by Aigner et al. [7] shows that visual analysis is usually studied within the information visualization area with a focus more on the visualization and exploration capabilities of the reviewed methods, and less on the pure analysis of the patients with a focus towards the discovery of new knowledge. As part of the shift in emphasis from *looking* at time-oriented data to *understanding* these data, recent visual exploration systems in general, and in the medical domain in particular, include capabilities for sophisticated interactive exploration of multiple-patients data set [8,9], a shift that is manifested also in recent *Visual Analytics in Healthcare* (VAHC) workshops [10-12]. However, these studies support only the analysis of *raw* data, and do not focus on the necessary underlying *knowledge* required for a more sophisticated analysis.

At the other end of the analytical spectrum, one finds the pure computational *data mining* (DM) methods (in particular, *temporal data mining* (TDM) methods), i.e., the data-driven approaches. These methods are automated, computationally valid, and complete (i.e., discover all patterns in the data); but most of them are not interactive, are intended only for a "super-user" with significant experience, and do not allow an effective exploration of the (typically too numerous) computed output, much of which is irrelevant. As result, the small number of significant results might be missed. Thus, data-driven methods are *complete*, but their *precision* is low, in the sense that most of the discovered patterns are of low significance.

A recent group of algorithms attempts to enumerate automatically, in a data-driven fashion, all frequent (i.e., above a given threshold of support) temporal patterns, given a set of symbolic time intervals based on temporal abstraction (TA) methods [13, 14]. These patterns are referred to as *Time-Interval Related Patterns* (TIRPs). A comprehensive description of the development of time intervals mining is in [15], including a detailed description of the recent KarmaLego framework, which has been demonstrated as significantly faster than previous TIRP-discovery approaches, due to its efficient data structures and its exploitation of the transitivity properties of temporal relations [15,16].

Therefore, we had designed and implemented a framework that combines both types of application, i.e., query-driven and data driven. We call our overall framework, which integrates several different computational models, and visual tools, the Visual Temporal Analysis Laboratory (ViTA-Lab). Recently, such a combination of visual and analytical capabilities of data analysis has been referred as *visual analytics* [17]. Thus, one might refer to the ViTA-Lab framework as a type of a visual analytics system.

2. The key principles of the ViTA-Lab framework

- To significantly enhance the capabilities of a visual analytics system, we propose to first preprocess the input raw, time-stamped data to produce a set of clinically meaningful summarizations and interpretations, typically interval-based, known as *temporal abstractions* (TAs) (or abstract concepts). In our framework, we use the *knowledge-based temporal-abstraction* (KBTA) method [18]. Using a domain-specific temporal-abstraction knowledge base acquired from a clinical domain expert, this method derives context-sensitive temporal abstractions from raw time-stamped data; for example, given a series of time-stamped hematological concepts, such as white blood-cell counts, and a series of time-stamped raw liver-enzyme values, the pattern, "a period of more than two months of grade II or higher bone-marrow toxicity, followed within three months by a period of at least one month of decreasing liver-functions" might be derived. (In this case, the medical concepts (terms) such as "grade II of bone-marrow toxicity" and "decreasing liver-functions" are defined within a particular clinical context; for example, a particular oncology protocol).
- The temporal data can be analyzed in parallel either by using the data-driven (data–mining) computational methods, which automatically discover frequent temporal patterns, or by using the interactive visual exploration interface to explore the data (both raw data and TAs) of multiple patients, and to understand the inter-relationships among concepts over time.
- The ViTA-Lab framework includes several visual interfaces:

  1) The main visualization and exploration interface provides an interactive overview of the raw longitudinal concepts and of the distribution of the derived temporal abstractions for multiple patients, at any temporal granularity. This interface is based on the VISITORS framework [19]. The VISITORS system enables the user to display on the fly multiple TAs for a patient population and interactively explore their distributions over time, zooming in and out of the timeline at multiple temporal granularities. When the user selects a raw or abstracted (derived) concept in the knowledge browser, the concept is displayed in a panel on the right.

  2) The Temporal Association Chart (shown below in section 3) enables the user to visually display probabilistic temporal associations among the distributions of multiple different concepts at different times. This interface is based on the parallel coordinates paradigm [20] to show relationship between attributes, i.e., concepts. The full description of the temporal association chart functionality and analysis capabilities can be found elsewhere [21].

  3) The Patterns Explorer (also shown below in section 3) is a dedicated interface for data-driven mining of frequent patterns in the data and for their exploration. Its underlying semantics are based on a version of the KarmaLego algorithm for the discovery of frequent temporal patterns [15].

3. A Proof of Concept in the Diabetic-Patients Domain

To demonstrate the potential benefits of the new framework, we introduce below a real scenario demonstrating the discovery of new patterns and the analysis of time-oriented data. The scenario explores the data of a group of 1711 diabetes patients who had been followed for at least five years. Our focus will be on determination of the best medication for achieving the important objective of reducing the HbA1c (glycated hemoglobin) measure (the main monitoring method for diabetic patients, which represents the mean blood-glucose level of the patient over the past three months). The medications to be investigated will be two popular oral anti-diabetic agents: Metformin (a drug from the biguanid class, which reduces blood glucose through the suppression of glucose production by the liver); and Glibenclamide (a drug from the sulfonylurea class, which causes the pancreas to secrete more insulin). Metformin and Glibenclamide are the only two oral anti-diabetics in the World Health Organization Model List of Essential Medicines.

To create a small knowledge base, the HbA1c levels (grade_1 up to grade_4) were specified by the medical domain expert, as represented in Table 1. Since there are no agreed symbolic ranges with respect to the drug doses (daily), the levels of the medications were computed by the Equal-Frequency discretization method from the data of all patients (see Table 2).

Table 1. The predefined symbolic HBA1C levels

HbA1c_State Levels	HbA1c values
grade_1	<7
grade_2	7-9
grade_3	9-10.5
grade_4	>10.5

Table 2. The predefined symbolic medication levels

Metformin_ State Levels	Metformin values (mg)	Glibenclamide_ State Levels	Glibenclamide values (mg)
grade_1	<850	grade_1	<15
grade_2	850-1700	grade_2	15-30
grade_3	>1700	grade_3	>30

4.1. A First Overview of the Patients' Data

Figure 1 presents an overview of the HbA1C distribution over time, using a Temporal Association Chart (TAC), a very particular type of an exploration operator (a complete description of the computational methods underlying TACs and their formal semantics appears elsewhere [20]), of the HbA1c_State abstract concept across the first five years of therapy for a group of close to 1700 patients (1300 patients annually, on average; not all patients have data values in all time windows). (Note that the patient data are aligned not by their absolute date, but according to the start of treatment, i.e., a relative time line operator is applied, using as a reference point for time zero the start of therapy for each patient.)

The distribution of HbA1c levels (the color legend of values is denoted by "2"), i.e., the relative proportion of patients who have each HbA1c state in each year of follow up, is represented within the specific time interval (denoted by "3") according to an annual (1 year) granularity view (as represented in panel denoted by "1"). Thus, for example, the annual HbA1c state for 33.79% of the patients in the first year of follow up was summarized (using a delegation function) as "grade_2" (see the tooltip denoted by "4").

The red links between two values of different adjacent distributions represent a group of patients who had both of the values, and thus an association over time between these values: The deeper the shade of the link's color, the higher the level of confidence in the relationship. That is, a darker shade of red indicates a higher confidence level in the future (right) value, given the current (left) value. The width of the link corresponds to the level of support, i.e., number of subjects having this particular value combination for these two time windows; broader links represent an association with higher support. By hovering with the mouse over an edge, the user will see additional statistical information, such as the lift and the statistical proportional test (see yellow tooltip denoted by "5"). In this case, the tooltip represents the relation between the "grade_1" values of the HbA1c_State during the third and the fourth year of treatment: 27% of the patients in the relevant patient group had this particular combination of values (i.e., support = 0.27), and 68% of the patients with "grade_1" HbA1c_State values during the third year also had "grade_1" HbA1c_State values during the fourth year (i.e. confidence = 0.68). This temporal association was valid for 461 patients, with a lift of 1.34. A proportion test that compares the confidence of the association (68%) to the prior probability of the HbA1c_State having the value grade_1 in the fourth year (51%), using the actual patient numbers, was significant with $p<0.05$ ("True"). Thus, this is a significant, non-random temporal association.

Three main "clusters" or "pathways" could be visually identified in Figure 1:

1) The main cluster of patients with the "grade_1" value of HBA1C_State (denoted by "6"), i.e., although in the case of nearly 100 patients the HbA1c levels changed from year to year, most of the patients each year are balanced (with respect to the results of their therapy), with low values of HbA1c. Support values across the 5 years range approximately around 26% to 27%, with confidence of 65% to 74%. In other words, a third of the investigated patients are stable and managed well, with respect to HbA1c values.

2) Similarly, we identify the group of nearly 300 patients with a "grade_2" value of HbA1c_State over the five years (denoted by "7"). (These are not necessarily the same patients every year, but a clear mass across the temporal pathway can be observed).

3) Finally, we note a segment of the patient population who are not balanced, having one of the higher HbA1c_State values (denoted by "8"); here the support and confidence of the relationship between the values change over the years.

In this scenario, we would like to focus only on patients who are at the greatest level of risk for developing complications of diabetes: those who have had the "grade_4" value of HbA1c_State during the first year of the treatment (the rest of patients have been filtered out), as displayed in Figure 2.

4.2. A Detailed View of the Patients who have had High HbA1c Levels

Figure 2 shows interrelations between the HbA1c_State values from the second and following years for patients who have had a "grade_4" value for that state abstraction during the first year. As we can see from screen shots, the behavior over time of the patients actually varies quite a bit. The initial group of 142 patients has different pathways: the state of HbA1c for some of the patients is improved (i.e., changed to a lower level, from "grade_1" to "grade_3") while another group maintains the high-risk, "grade_4" value of the HbA1c_State.

Of course, patients move among the groups, and the change might be mediated by external factors that are not shown in the TAC, such as different types of medications. Since visually it is very difficult to recognize certain patterns, the user will now apply the data-driven, computational TDM engine, with the goal being the discovery of temporal patterns that might characterize these two different groups of patients (i.e., those whose state was improved to a lower HbA1c level, and those whose state remained at the risky, high level).

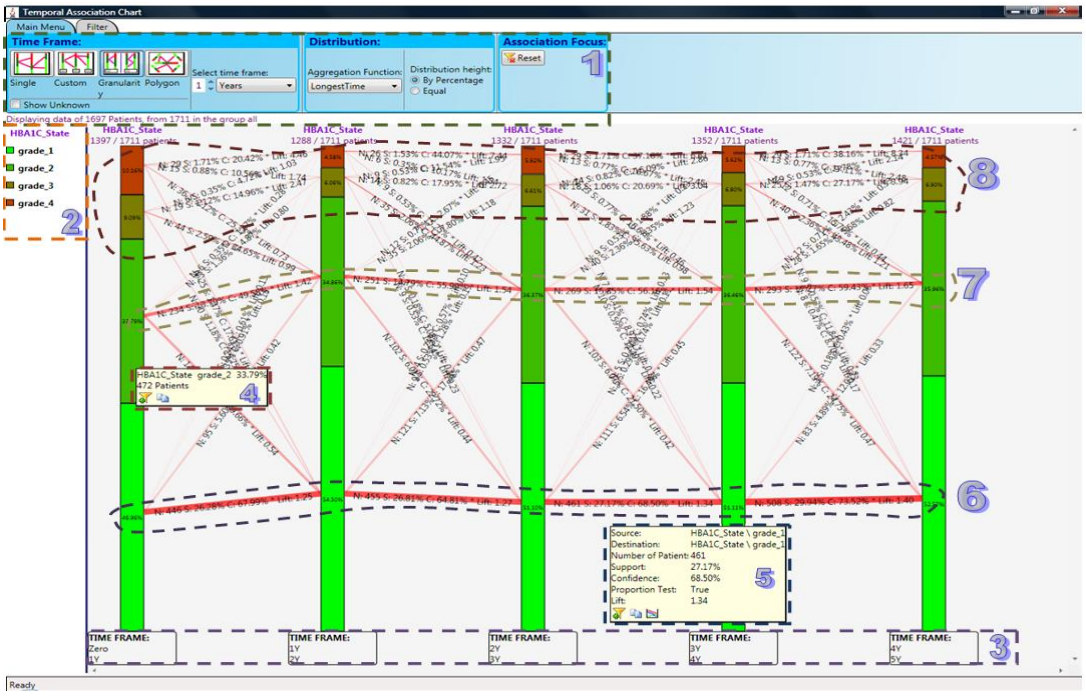


Figure 1. A Temporal Association Chart, showing the interrelation amongst HbA1c levels across the first five years after start of therapy. In a TAC, subjects who have both of the values of two temporally adjacent distributions are connected by a line; broader lines represent greater support for that conjunction of values, while a deeper shade of the linking color (red, in this case) represents greater confidence (i.e., the probability of having the second [right] value given the first [left] one). For example, the lowest temporal path (denoted by "6") represents a cluster of patients who have had Grade-1 HbA1c values over all of the five years. Note that the timeline is a relative one, aligned such that time zero is the start of the therapy for each patient. See text for complete description.

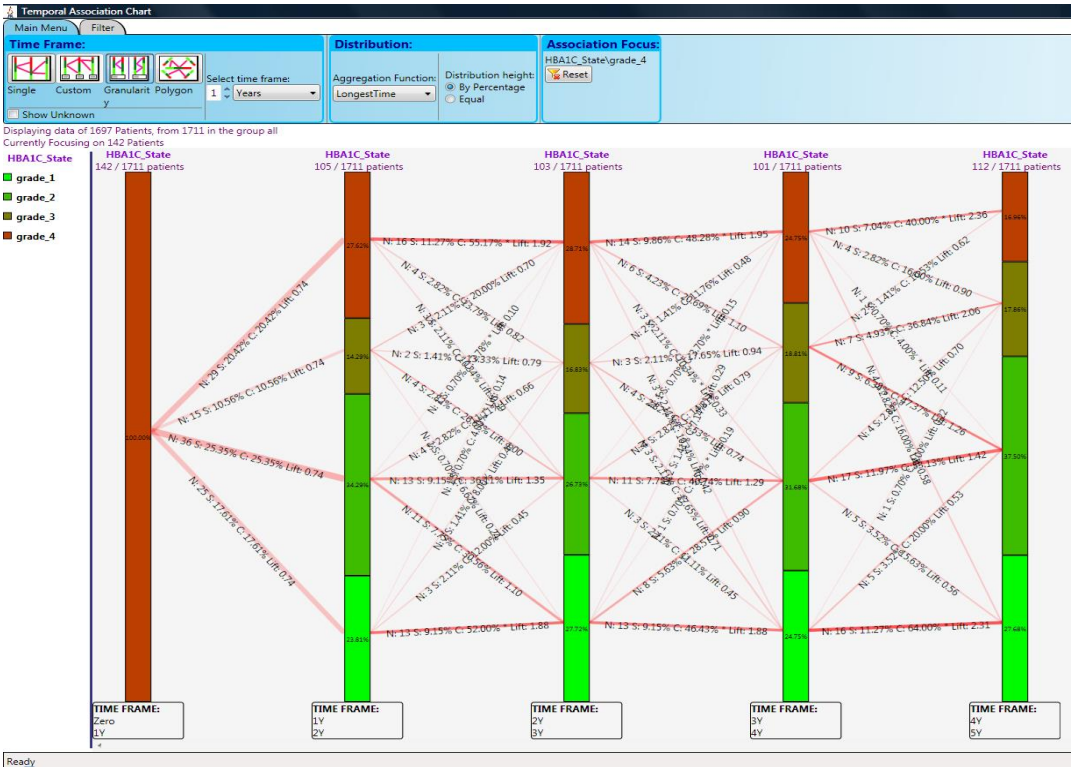


Figure 2. A Temporal Association Chart showing the interrelation of HbA1c across the first five years after start treatment, displayed only for patients who have had a "grade_4" value of HbA1c_State during the first year of therapy.

4.3. Temporal Patterns of Patients with High HbA1c Levels

To clarify the HbA1c state of the patients that improved and to try and associate it with the various types of medication used (from a list of more than 10 medications that were used by this particular population), the user now applies the KarmaLego computational engine, requesting to discover only frequent patterns of various HbA1c_State values and any medications.

Applying the data-driven TDM mode returns 439 different patterns that involved the HBA1C_State and medication-level state abstractions, given a minimal support level of 10% (i.e., at least 10% of the patients had to have at least one instance of that pattern). By using the filtering engine that we added to the TDM engine (denoted by the label "1" in Figure 3), we are focusing on temporal patterns that starting with a “grade_4” value of the HbA1c_State and include in the rest of the pattern any HbA1c_State values. Only two medications (Metformin and Glibenclamide) passed the 10% support threshold. Patterns in which Metformin and Glibenclamide treatments were involved are labeled as "2" (panels "a" to "d") and "3" (panels "e", "f"), correspondingly.

For instance, panel "a" shows the temporal pattern in which a period of medication treatment by Metformin (the middle long interval) with a dose level of “grade_2,” followed the “grade_4” period of HbA1c_State (denoted as a short interval on the left). The long interval on the right represents the value “grade_3” of the HbA1c_State. By the temporal position and duration of intervals, the user can easily understand various temporal relations among the key concepts from which the patterns are derived (e.g., before, overlaps, meets). For convenience, the values legend appears on the left side in the panel. The “VS” symbol denotes the vertical support value of the pattern (i.e., portion of subjects in which it occurs at least once), e.g., 15% of patients have had pattern shown in panel "a". (A horizontal support level denotes the mean frequency of the pattern within each subject in which it was found at least once).

The user is able also to examine the mean duration and mean time gap between pairs of intervals and the relevant temporal relation among them (as shown in panel "d").

It is clear that there is a frequent pattern of using one of the two most common medications and reducing the HbA1c level.

One could ask *which medication is associated with a higher rate of improvement in the HbA1c_State values*?

In fact, when one examines all patterns, overall, throughout all therapy years, it turns out that the HbA1c state of 77 out of 112 patients who had HbA1c_State of "grade_4" before the medication period, and had taken only Metformin, seemed to be improve to a lower level (as shown in panels "a" to "d"); the mean duration from the start of the medication interval to the start of the lower HbA1c_State value interval was 9 months. To obtain the information regarding the number of subjects in each pattern, and, especially, regarding the number of distinct (or common) subjects, given several selected patterns, we developed an additional form (not shown here).

Conversely, throughout all years, 34 out of 86 patients who had HbA1c_State "grade_4" and who had taken only Glibenclamide (panels "e", "f") (although no frequent pattern that ended with a "grade_1" value was discovered in that group); the mean duration from the start of the medication interval to the start of the lower HbA1c_State value was 14 months. 34 patients were treated by both medications. The difference, using a proportion test, among the improvement rates (with respect to HbA1c values only) of the two medications would usually be considered as quite significant ($p = 0.04 < 0.05$).

To sum up, as a result of the application of the TDM engine, and through further exploration of temporal patterns, we identified two general clearly distinct groups of patients: (1) those in whom the high first year HbA1c_State seems to be improved by medications (78 patients), and (2) those in whom the HbA1c_State seems to remain high (31 patients). (The other 33 patients of the original 142 were omitted from the analysis, due to missing data during other years, etc.)

Furthermore, it seems that in this particular group of patients, Metformin was associated with a significantly higher rate of decreased HbA1c_State (as well as with a shorter interval of achieving that decrease) than Glibenclamide.

Conclusions

We have presented an advanced, iterative, visual analytical framework, integrating data-driven and user-driven analysis of time-oriented clinical data, and capitalizing on a knowledge-based temporal-abstraction preprocessing phase. The fast intuitions provided by such a framework can often be translated into a deeper analysis or further studies.

Thus, we would suggest that the ViTA-Lab framework might potentially serve as a virtual laboratory for clinical investigations of large masses of longitudinal clinical databases.

Acknowledgements

We would like to thank Prof. Avi Porath for making accessible the clinical data used in this research. Part of this research was supported through the EU 7th Framework project, MobiGuide, EU award No. 287811.

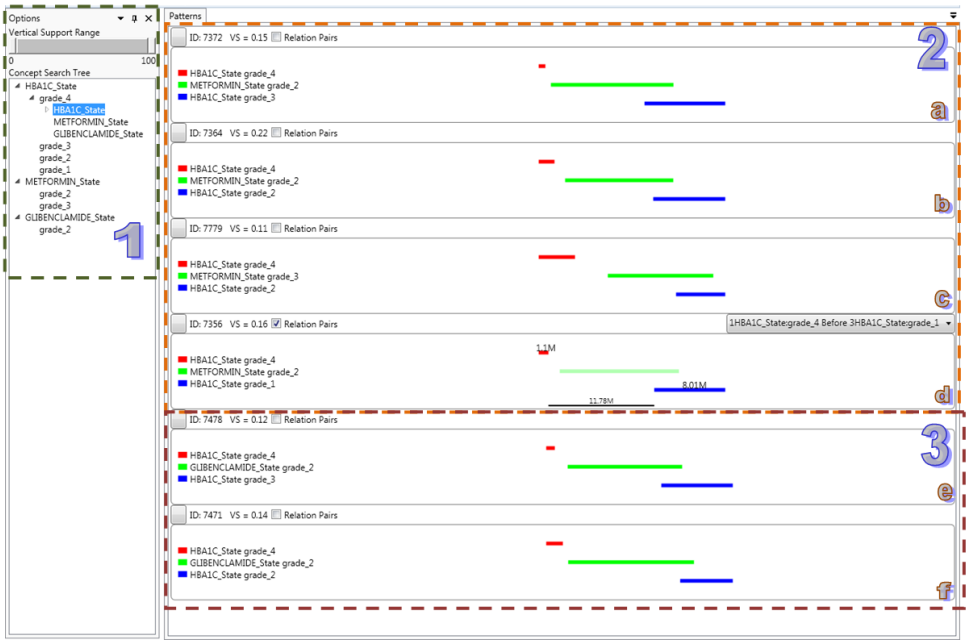


Figure 3. Patterns for the improving (i.e., decreased HbA1c state values) group of patients

References

[1] Plaisant C., Mushlin R., Snyder A., Li J., Heller D. and Shneiderman B. (1998) LifeLines: Using visualization to enhance navigation and analysis of patient records. American Medical Informatic Association Annual Fall Symposium (1998), pp. 76-80,

[2] Wang T., Plaisant C., Quinn A., Stanchak R., Shneiderman B. and Murphy S. Aligning temporal data by sentinel events: Discovering patterns in electronic health records. SIGCHI Conference on Human Factors in Computing Systems 2008.

[3] Falkman G. Information visualisation in clinical Odontology: multidimensional analysis and interactive data exploration, Artificial Intelligence in Medicine, vol. 22(2), pp. 133-158, 2001.

[4] Chittaro L., Combi C. and Trapasso G. Visual Data Mining of Clinical Databases: An Application to the Hemodialytic Treatment based on 3D Interactive Bar Charts. Proceedings of Visual Data Mining VDM'2002, Helsinki, Finland, 2002.

[5] Chittaro L. Information visualization and its application to medicine. Art Intell Med 2001; 22(2):81-88.

[6] Rind A, Wang T, Aigner W, Miksch S, Wongsuphasawat K, Plaisant C, Shneiderman B. Interactive Information Visualization for Exploring and Querying Electronic Health Records: A Systematic Review. Foundations and Trends in HCI, 5(3):207-298. 2013.

[7] Aigner W., Miksch S., Müller W., Schumann H. and Tominski C. Visual Methods for Analyzing Time-Oriented Data. IEEE Transactions on Visualization and Computer Graphics 2008; 14(1): 47-60.

[8] Wongsuphasawat K, Alexis Guerra Gómez J, Plaisant C, David Wang T, Taieb-Maimon M and Shneiderman B. LifeFlow: Visualizing an Overview of Event Sequences in Proceedings of the 2011 Annual CHI'11 Conference : 1747-1756.

[9] Zhang .Z Ahmed F, Mittal A, Ramakrishnan IV, Zhao R, Viccellio A, Mueller K. AnamneVis: A Framework for the Visualization of Patient History and Medical Diagnostics Chains, " IEEE VisWeek Workshop", 2011.

[10] Zhang Z, Gotz D and Perer A. Interactive Visual Patient Cohort Analysis. VAHC-2012, J. Caban, D. Gotz (ed.). 2012.

[11] Brodbeck D, Degen M and Walter A.Masterplan: A Different View on Electronic Health Records. Proceedings of the IEEE VisWeek Workshop on Visual Analytics in Healthcare – VAHC-2012, J. Caban, D. Gotz (ed.). 2012

[12] Meyer T, Monroe M, Plaisant C, Lan R, Wongsuphasawat K, Coster T, Gold S, Millstein J and Shneiderman B. Visualizing Patterns of Drug Prescriptions with EventFlow: A Pilot Study of Asthma Medications in the Military Health System. VAHC 2013.

[13] Moskovitch R, Shahar Y, Classification Driven Temporal Discretization of Multivariate Time Series, Data Mining and Knowledge Discovery, DOI: 10.1007/s10618-014-0380-z, 2014.

[14] Höppner F. Time Series Abstraction Methods -- A Survey. In Proceedings GI Jahrestagung Informatik, Workshop on Knowl. Discovery in Databases, LNI, Dortmund, Germany, 777-786, 2002.

[15] Moskovitch R. and Shahar Y. Fast Time Intervals Mining using the Transitivity of Temporal Relations. Knowledge and Information Systems. Springer-Verlag, 2013. doi:10.1007/s10115-013-0707-x

[16] Moskovitch R. and Shahar Y. Classification of Multivariate Time Series via Temporal Abstraction and Time Intervals Mining, Knowledge and Information Systems, In Press, 2014.

[17] Keim D., Mansmann F., Oelke D. and Ziegler H. Visual Analytics: Combining Automated Discovery with Interactive Visualizations. Discovery Science,pp. 2-14, 2008.

[18] Shahar Y. A framework for knowledge-based temporal abstraction. Artificial Intelligence, 90(1-2):79-133, 1997.

[19] Klimov D., Shahar Y and Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. Artificial Intelligence in Medicine, 49(1):11-31. May 2010.

[20] Inselberg A. Parallel Coordinates: Visual Multidimensional Geometry and its Applications. Springer, 2009.

[21] Klimov D., Shahar Y. and Taieb-Maimon M. Intelligent interactive visual exploration of temporal associations among multiple time-oriented patient records. Methods Inform Med 2009; 48(3):254–62.

Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels

Eugenia McPeek Hinz[1], David Borland[2], Hina Shah[3], Vivian L. West[3], W. Ed Hammond[3]
[1]Duke Health Technology Solutions, Duke University; [2]RENCI, The University of North Carolina at Chapel Hill;
[3]Duke Center for Health Informatics, Duke University

Abstract

Diabetes mellitus is a chronic long-term disease requiring consistent medical treatment to achieve glucose control and prevent complications. Time of diabetes diagnosis can be variable and delayed years beyond disease onset. The spectrum of glycemic trajectories for a general population over an entire diabetes disease course is not well defined. Aligning disease course on death enables coherent data visualization. Our temporal visualization tool uses a parallel-sets inspired technique that illustrates the complicated and varied trajectories of hemoglobin A1c levels for a general diabetic population. A consistent glucose normalization trend for the majority of patients is seen over the course of their disease, especially in the six months prior to death. This tool permits discovery of population-level Hemoglobin A1c trends not otherwise evident without disease phase synchronization. These findings warrant further investigation and clinical correlation. Visualizations such as this could potentially be applied to other chronic diseases and spur further discoveries.

Introduction

Diabetes mellitus is a chronic disease that affects millions worldwide, resulting in numerous cardiovascular and renal complications, and subsequently is a major cause of death. Age of onset, duration of diabetes, and poor glycemic control are well-defined risk factors for the development of complications associated with increased mortality in persons with diabetes mellitus.[1] To decrease the development of complications associated with diabetes, tightly controlled glucose is the standard of care.[2] Notably some large prospective trials have found either worse outcomes or lack of benefit for some patients at high risk for complications under tight treatment control regimens.[3,4] Hemoglobin A1c (HbA1c), a marker of glucose control over the two to three months preceding the test, is a validated predictor of diabetes-related complications.[2] Using HbA1c to understand trajectories and temporal patterns of glycemic control over an entire diabetes disease course could be an important factor in improving treatment and reducing overall complications.

Data visualization techniques offer opportunities to explore large datasets and identify clinical patterns that might otherwise not be obvious. In this study we present a cohort of patients with diabetes (via ICD9 codes) from Duke University's data warehouse, visualizing their HbA1c levels over time, aligned by death, to explore trends of glycemic control. To the best of our knowledge, temporal visualization of glycemic control for a diabetic population standardized on death has not previously been presented. Our visualization groups HbA1c values into ordered categories of glycemic control (Normal, Borderline, Controlled, and Uncontrolled), utilizing a method based on parallel sets[5] and Sankey diagrams[6] to view temporal patterns in HbA1c values. We incorporate a number of features to facilitate interactive data exploration, such as viewing the progression of values either forwards or backwards in time, the ability to change the temporal sampling and range of the data being viewed, highlighting of multiple subpopulations, coloring based on the category along each path in the data or at the beginning/end of each path, and the incorporation of demographic data, such as gender.

Related Work

Analysis of diabetes indicators

A reduction in HbA1c levels lowers the risk of diabetes-related complications and mortality, especially for patients earlier in their disease course.[7] Counterintuitively, intensive treatment of glucose to reach near-normal levels for patients already experiencing diabetes-related complications has failed to lower all-cause mortality.[3] While large cross-sectional studies of populations such as the National Health and Nutrition Examination Survey find a temporal trend toward improving glycemic control over time, less well-established is the temporal trajectory of glycemic control for diabetic patients in general.[8] The only other work the authors are aware of looking specifically at glycemic control trajectories for a large diabetic cohort followed patients prospectively to the end point of death.[9] The study correlated initial glucose control to outcome of death, but did not report specifically on the population glucose trajectories.

Visualization methods

Our visualization tool is based on parallel sets[5] and Sankey diagrams.[6] Parallel sets were originally developed for visualizing relationships in multivariate categorical data, whereas Sankey diagrams, introduced by M. H. P. R. Sankey, are typically used for describing the flow of quantities such as energy, material, or cost. The original parallel sets user interface enables user-defined classification definitions, statistical analysis information, and various sorting methods. Parallel sets combines the concepts of parallel coordinates[10] and mosaic plots[11], enabling an aggregation of data points within visualization elements, as opposed to showing each individual element, which is typical of parallel coordinates. Multiple systems aggregate data points for summary.[5,12-14] For example, EventFlow enables the search and visualization of interval data, such as periods of medication treatment, to examine the order of sequences of events in the data.[12] OutFlow facilitates analyses of temporal event data in the form of pathways with relevant statistics.[14] All of these visualization tools look at event occurrences and their order, without placing these events on time axes. Our diabetes visualization uses the parallel sets paradigm, with each axis representing a temporal sample of HbA1c levels instead of a separate variable, similar to von Landesberger et al.[13] Although our current dataset is relatively small (121 patients), we chose a parallel sets representation in part due to its ability to aggregate many data points. The visual complexity is bounded by the number of axes and categories per axis, not by the number of data points, making it suitable for the exploration of larger datasets in the future. This representation also easily incorporates additional non-temporal variables, such as demographic data.

Methods

Data extraction and preprocessing

Data from Duke University's data warehouse were extracted using DEDUCE, an on-line query tool developed at Duke to assist researchers in human subjects research and departments seeking quality improvement data.[15] Beginning with over 4.4 million patients, we first queried by 23 IDC9 codes for diabetes mellitus, with and without complications. The query was further refined by querying on patient death indicator and laboratory tests for glycosylated hemoglobin (HbA1c), and finally by including only patients prescribed anti-hyperglycemics. This search returned data from 208 patients. From this cohort of 208, we eliminated four that did not have a year of death recorded, one whose date of death was documented but continued to have laboratory results recorded after that date, and 82 who did not have at least 10 years of HbA1c laboratory values. Our final cohort includes data from 121 patients.

We average HbA1c values, given as a percentage of total hemoglobin, over 6 month time intervals. In the case of missing HbA1c values within a 6 month period we first attempt to impute an HbA1c value from the average glucose (AG) values over that period of time, via the formula $HbA1c = (AG + 46.7) / 28.7$.[16] If no glucose values exist in that time period, the previous HbA1c value (measured or imputed) is carried forward. HbA1c values are then classified into four categories based on the severity of diabetes: Normal < 5.7, Borderline $[5.7, 6.5)$, Controlled $[6.5, 8)$, and Uncontrolled ≥ 8.

The sampled data is time-aligned by the death event for each patient. The visual representation of diabetes progression propagates backwards in time initially. Time is represented as number of years before death.

Visualization

Our visualization tool was developed using the D3 Javascript library.[17] The aim of this visualization is to investigate temporal trajectories of HbA1c levels for a large cohort of diabetes patients over a number of years prior to death. Since parallel sets is effective for showing relations between categories using aggregated frequencies of paths through categories at each dimension, it is a reasonable choice for showing HbA1c summary trajectories. The visualization tool shows a total of five categories: four representing glycemic control, and one optional Missing category for patients with data greater than 10 years before death. . Each vertical axis is a time step. The user can choose the frequency of these time steps, with a minimum sampling frequency of six months. The user can also select the maximum number of years before death.

The death event axis is placed at the right with all other time steps moving backwards in time to the left (Figure 1). Each vertical axis is split into the four HgA1c categories (Normal in green, Borderline in blue, Controlled in orange, and Uncontrolled in red), and a Missing category in grey prior to 10 years before death. The height of each axis category represents the proportion of the patients in that category at that point in time. Paths moving between axes recursively split moving backwards from death to show the trajectories of similar groups of patients. The visualization can show trends either starting at the death event i.e. going backwards in time (dividing

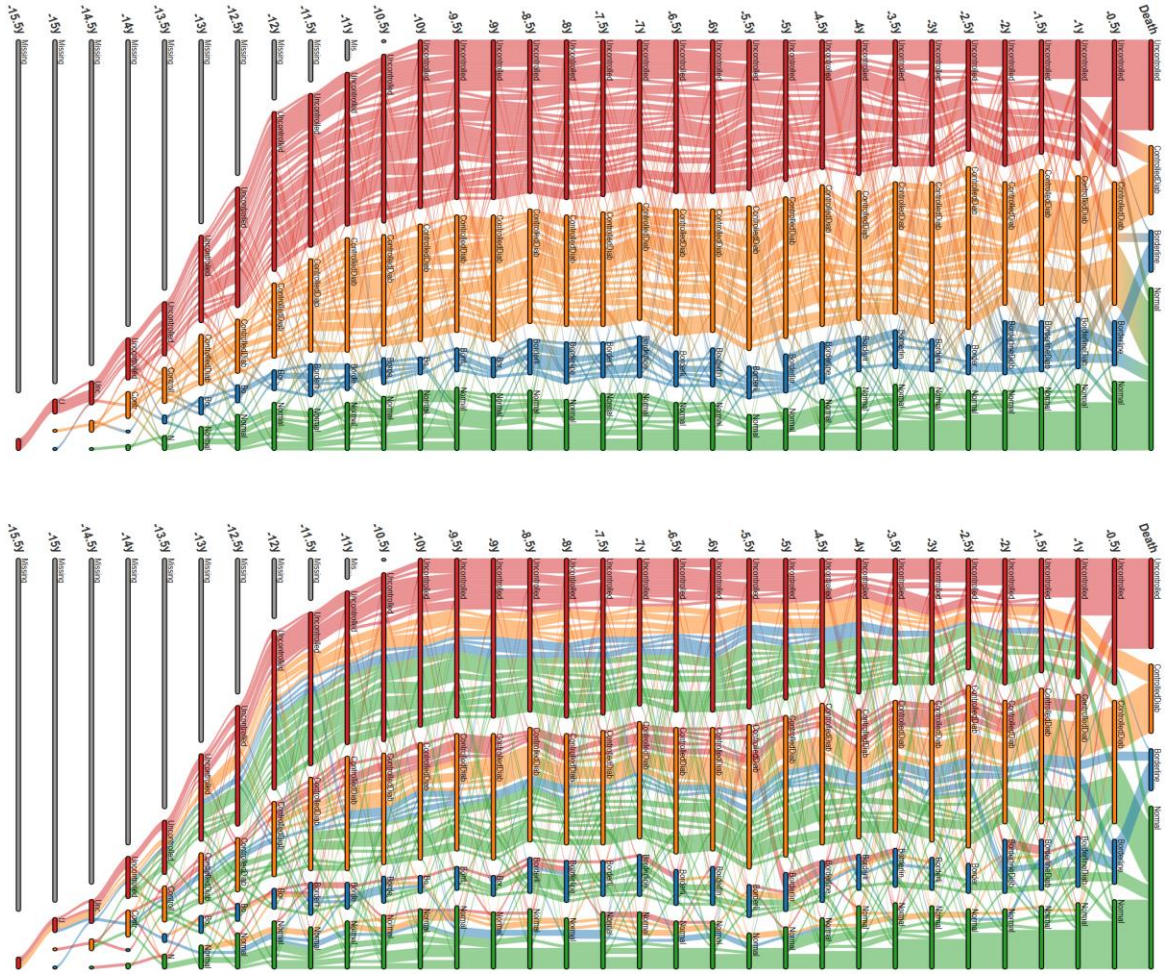


Figure 1. Diabetes progression overview visualizations. The top image colors paths by the current HbA1c at each time step, which is useful for emphasizing overall temporal trends. The bottom images colors paths by the HbA1c level at death, showing at each time step where each path will end.

recursively right to left), or starting at the last year in the visualization, i.e. going forward in time (recursive division from left to right). Going backwards and coloring by death shows at any time point the relationships between patients in a given category and their categories at death, while going forward in time shows the relationship between patients in a given category and their categories at a user-defined earlier point in time (Figure 2). Following Shneiderman's Mantra[18] of first overviewing and then filtering, the user can highlight one or more groups of patients by clicking on categories or trajectories to highlight the behavior of that group of patients going backward and forward in time, reducing visual clutter (Figure 2). A tooltip also shows the actual number of patients in each group and their percentage of the total population.

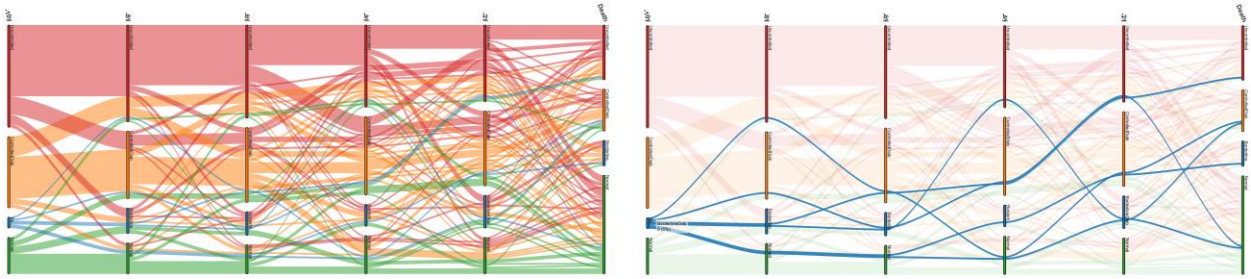


Figure 2. A 10-year range of data, sampled every two years, with forward propagation to show how the trajectories of patients change moving forward in time (left). Highlighting enables a focused view of a single category, reducing visual clutter (right).

The user can also chose between different types of coloring schemes for the paths: 1) color by the category at the first or last year (depending on the propagation direction), which shows the level of variation for a category over the length of the visualization, 2) color by transition, where the transition has a gradient from the source to target category color, which is useful for showing overall trends, and 3) color by reverse transition, where the transition path has a gradient from the target category to the source category, which is useful for category-level analysis of the distribution of source and target categories at a particular time step's category (Figure 3). To reduce visual clutter there is also an option to look at only static transitions (i.e. no change in category between time steps), and to look at only variations (i.e. only changes in the categories).

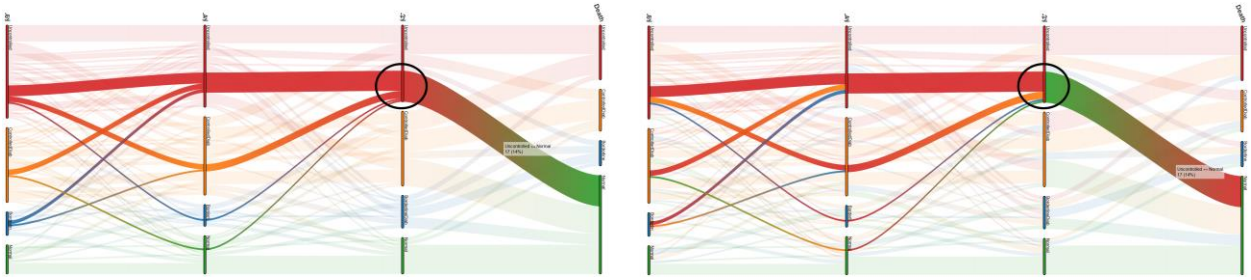


Figure 3. The user can observe separate groups by selecting individual trajectories. In addition to coloring by the starting category, paths can be colored by a gradient from source to target category (left), which redundantly encodes the category at each axis to emphasize overall trends, or by target to source category (right), which enables a rapid analysis of where paths are moving to/from at each category. The circled regions highlight this difference. On the right, it is immediately obvious what category this trajectory came from at death (Normal in green) and how this group is distributed at the previous time step.

We also include the ability to incorporate demographic data, such as gender, as additional axes (Figure 4). This feature enables the comparison of trajectories for different subpopulations based on data other than just HbA1c levels.

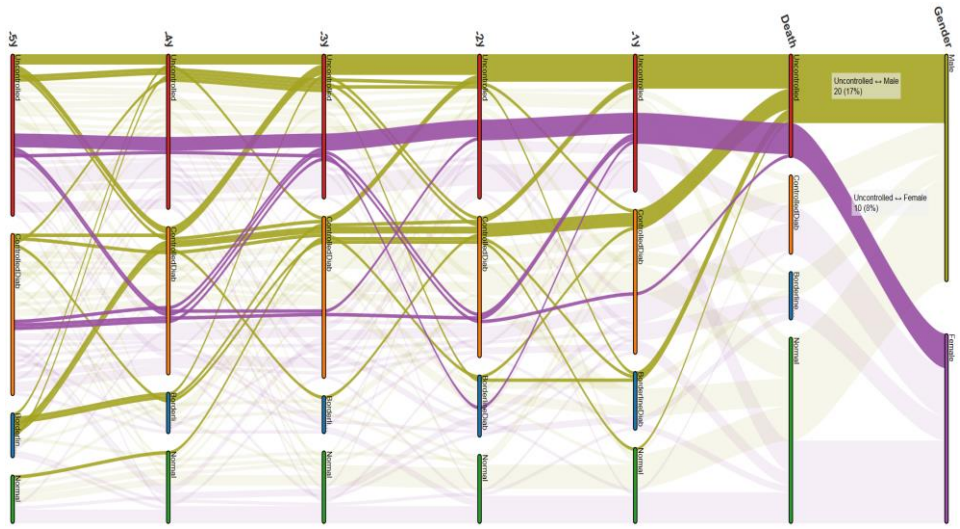


Figure 4. By adding a gender axis and selecting two groups we can compare the variability of males who were Uncontrolled at death (olive) to women who were Uncontrolled at death (purple). Men appear to have more variability over the 5-year period being visualized, as shown by the large number of transitions between different categories.

Findings

In the 10 years before death, there is a consolidating trend to improved glucose control across all diabetes control categories from uncontrolled to normal. Overall diabetes control shifts from uncontrolled diabetes for 46% of the cohort to 25% at death utilizing HbA1c and imputed glucose values. A reciprocal increase in combined borderline and normal range glucose control goes from 25% at 10 years out to 57% at death (Table 1). The trend for better glucose control is most visible in the last six months before death. The overall final glycemic trajectory is also evident in the bottom image from Figure 1 where the control category at death is colored retrospectively. Notably a

small minority of the uncontrolled sub-group remains poorly controlled over the entire disease course. By including category temporal transitions this visualization also illustrates the complexity of the underlying data, with many trajectories exhibiting a large degree of variation in HbA1c categorization over time.

Table 1. Percent of patients by Diabetes control category over 10 years prior to death using HbA1c with imputed glucose results.

Glycemic Control by HbA1c	*-10 years years to death*	*-5 years years to death*	*At Death*
Uncontrolled Diabetes	46 %	39 %	25 %
Controlled Diabetes	32 %	39 %	19 %
Borderline	5 %	11 %	12 %
Normal	17 %	12 %	45 %

Discussion

The progression of diabetes with accumulating end organ complications is well recognized. There is a clinical presumption that diabetes-related complications are also associated with worsening glycemic control for patients with end stage diabetes mellitus. Since most prospective cohort studies are organized by a patient's clinical presentation, treatment or demographics, they tend to be cross sectional studies of a population and include patients across a disease continuum. By creating a cohort organized by a death criterion with 10 or more years of diabetes lab data, we have sub-selected a general but presumably more ill diabetic population. Phasing HbA1c values by death allows data coherence that translates into the visualization of glycemic trajectories that would be less evident in cross sectional studies of diabetic patients. Understanding the course of diabetes control is important to discerning differences in outcomes, treatments and identifying sub-phenotype populations.

Death event as an organizing point for temporal data visualization permits a clear starting point to observe the course of medically treated diabetes. Cause of death is not defined, so further characterization of subpopulations visualized in the cohort, like the always uncontrolled diabetes subgroup, warrants further clinical investigation to see if they are representative of the cohort overall. All patients in this cohort had data for at least 10 years, as such our population is specific for patients under some manner of regular medical care, and interpretation of the data with respect to populations with less regular medical care should be limited. Using the imputed average glucose and average HbA1c values aligned on the cohort's endpoint enables capture of all glycemic values, including those potentially before even the diagnosis of diabetes is made.

We observed a trend to normalization of HbA1c in the last year of life. The reasons behind improved diabetes control near the end of life could include multiple factors, such as increased insulin half-life due to impaired renal and hepatic metabolism, decreased dietary intake related to anorexia or nausea, and falsely low HbA1c secondary to uremia or anemia.[19] Interestingly, the goals for end-of-life treatment in diabetic patients are generally to limit side effects of either hyper or hypoglycemia and often entail a scaling back of treatment which would be expected to be associated with more hyperglycemia not less. By using visualization tools to see the progression of HbA1c values in diabetic patients in the years before their death, our findings of glucose normalization in light of this paradigm highlight the need for further clinical investigation and interpretation.

Our data visualization tool displays temporal patterns of diabetes metric across a population and for the last years of this disease continuum. Tools such as these can only display patterns that can potentially illuminate findings that need further clinical validation and statistical investigation to determine clinical significance if any.

Future work

The visualizations we have shown here represent a small number of patients in the dataset. This has enabled us to test and refine the visualization before using large amounts of data. Next we will include diabetes-related co-morbidities, e.g. cardiovascular, neurological, and renal manifestations of prolonged diabetes illness, and additional demographic variables, e.g. age and ethnicity. We plan to link this temporal visualization to other multivariate visualizations highlighting selected groups of patients, helping to show factors related to diabetes. We are also working toward a better statistical analysis of the data, and its representation in this tool. In particular, we wish to incorporate information regarding the amount of imputed and extrapolated data in the visualization.

Conclusion

Exploring the natural disease course of diabetes control with data visualization tools permits identification of potentially clinically important trends that would be difficult to recognize otherwise. Further investigation and definition on the clinical significance of the normalization of HbA1c in the final years of life are warranted.

Acknowledgments
This work is supported 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. We acknowledge the assistance from Meghana Ganapathiraju in helping to refine the visualization. Our implementation was adapted from the d3.parsets reusable chart by Jason Davies. We also acknowledge the assistance of Mark Massing MD PhD MPH and Susan Spratt MD for comments and discussions on graphical representations of diabetes control.

References

1. The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med.1993;329(14):977-986.
2. American Diabetes Association. Standards of Medical Care in Diabetes. Diabetes Care. 2009;32(S1):S13-S61.
3. The Accord Study Group. Long-term effects of intensive glucose lowering on cardiovascular outcomes. N Engl J Med. 2011;364(9):818-828.
4. The UK Prospective Diabetes Study Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet. 1998;352(9131):837-853.
5. Bendix F, Kosara R, Hauser H. Parallel sets: visual analysis of categorical data. IEEE Symp on Info Vis. 2006;12(4):133-140.
6. Riehmann P, Hanfler M, Froehlich B. Interactive Sankey diagrams. IEEE Symp on Info Vis. 2005;233-240.
7. Holman R, Paul S, Bethel M, Matthews D, Neil H. 10-year follow-up of intensive glucose control in type 2 diabetes. N Engl J Med. 2008;359(15):1577-1589.
8. Ford E, Li C, Little R, Mokdad A. Trends in A1C concentrations among U.S. adults with diagnosed diabetes from 1999 to 2004. Diabetes Care. 2008;31(1):102-104.
9. Gebregziabher M, Egede LE, Lynch CP, Echols C, Zhao Y. Effect of trajectories of glycemic control on mortality in type 2 diabetes: a semiparametric joint modeling approach. Am J Epidemiol. 2010;171(10):1090-1098.
10. Inselberg A, Dimsdale B. Parallel coordinates. Human-Machine Interactive Systems. 1991;199-233.
11. Hoffman H. Exploring categorical data: Interactive mosaic plots. Metrika. 2000;51(1):11-26.
12. Monroe M, Wongsuphasawat K, Plaisant C, Shneiderman B, Millstein J, Gold S. Exploring point and interval event patterns: Display methods and interactive visual query. HCIL Tech Report, University of Maryland. 2012.
13. von Landesberger T, Bremm S, Andrienko N, Andrienko G, Tekusova M. Visual analytics methods for categoric spatio-temporal data. IEEE Conf on Vis Anal Sci and Tech(VAST) 2012;183(192):14-19.
14. Wongsuphasawat K, Gotz D. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. IEEE TransVis Comput Graph, 2012;18(12):2659-2668.
15. Horvath MM, Winfield S, Evans S, Slopek S, Shang H, Ferranti J. The DEDUCE guided query tool: Providing simplified access to clinical data for research and quality improvement. J Biomed Info. 2011;2:266-276.
16. Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ. Translating the A1C Assay into estimated average glucose values. Diabetes Care. 2008;31(8):1473-1478.
17. Bostock M, Ogievetsky V, Heer J. D^3: Data-driven documents. IEEE Trans Visualization & Comp Graphics. 2011;17(2):2301-2309.
18. Shneiderman B. The eyes have it: A task by data type taxonomy for information visualization. Proc. 1996 IEEE Symp Vis Lang. 1996;336-343.
19. Kalantar-Zadeh K, Derose SF, Nicholas S, Benner D, Sharma K, Kovesdy CP. Burnt-out diabetes: Impact of chronic kidney disease progression on the natural course of diabetes mellitus. J Renal Nutrition. 2009:19(1):33-37.

Detecting Novel Associations for Surgical Hospital Readmissions in Large Datasets by Interactive Visual Analytics

Stein Olav Skrøvseth, PhD[1,2], Adam Perer, PhD[2], Conor P. Delaney, MD, PhD[3], Arthur Revhaug, MD, PhD[4,5], Rolv-Ole Lindsetmo, MD, PhD[4,5], Knut Magne Augestad, MD, PhD[1,3,4]

[1] Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, Tromsø, Norway. [2] IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA. [3] Department of Surgery, University Hospitals Case Medical Center, Cleveland, Ohio, USA. [4] Department of Gastrointestinal Surgery, University Hospital of North Norway, Tromsø, Norway. [5] Department of Clinical Medicine, University of Tromsø, Tromsø, Norway

Abstract

Complete Electronic Health Record (EHR) information was defined for 4307 patients undergoing gastrointestinal surgery at a Norwegian university hospital. Outcome was readmission within 30 days of the index episode. All structured codes were extracted along with variables extracted from other parts of the EHR, including from free text, and included in an interactive visualization. The visualization provided clinicians with an interactive way to explore the data. We were able to identify a set of clinically important variables that strongly correlated positively or negatively with readmission. The variable most strongly associated with readmission was whether the length of the index stay was longer than 5 days. The patterns that were identified largely corresponded to clinical and published knowledge regarding hospital readmissions. Inclusion of multiple factors in the visualization permits confirmation of expected causes of readmission, and exploration of new and unknown patterns. The visualization tool gave a way to identify patterns relating to a clinically important outcome, and provided an interactive and meaningful way for clinicians to engage with their data.

Introduction

The Electronic Health Records of modern hospitals amass large amounts of data as part of routine care, and the use of such data for quality control and improvement of care should be a central tenet for all institutions where such data are available.[1] Due to the large amounts and veracity of the data, it is often cumbersome and complicated to manage, organize and test clinical hypotheses. However, modern analytics and visualization techniques provide novel ways in which users can analyze, categorize and organize such data enabling them to explore new hypotheses based on their datasets.[2]

Surgery is a clinical setting where the utility of visualization techniques on large datasets has been largely unexplored. Furthermore, critical complications and severe outcomes for patients are relatively common, and a large number of variables can impact given outcomes. Visual analytics of care pathways and patterns in patient characteristics is a way for clinicians and other healthcare providers can learn which surgical approaches lead to better or worse patient outcomes.

In the present paper, we focus on a clinically relevant problem: 30-day hospital readmissions as an example case study where visualization can provide clinically meaningful information. The Centers for Medicare & Medicaid Services (CMS) recently began using readmission rates as a quality metric and may lower reimbursement to US hospitals with excess risk-standardized readmission rates. This decision has lead to a huge interest on trying to identify patient factors associated with readmissions.[3,4] In particular, the research on readmissions within the surgical field has focused on identifying risk factors associated with preventable readmissions. Several risk factors have been identified, and it is shown that readmitted patients have distinct demographic and outcome variables. It is also shown that readmitted patients are older, have more comorbidities, longer operative times, and length of stay. Independent readmission predictors are higher American Society of Anesthesiologists score (ASA score), previous abdominal operation and intensive care unit stay.[5–8]

Building personalized risk assessment models for these patients is an important task, but such models can only use what is actively inserted into them, potentially missing important information. By non-

discriminately inserting all available data into the analytics, novel and possibly unsuspected patterns can be revealed visually and subsequently this information can be built into predictive models. However, these visualization methods are largely hypothesis generating and need validation by independent data.[9] We sought to investigate whether we could identify similar risk factors in a large surgical dataset, using novel visual analytics techniques on EHR data.

In the current manuscript we demonstrate a tool to visualize the information in an EHR to identify the most important variables associated with a clinically relevant question, i.e. variables associated with 30-day readmission in a surgical population. In particular we used information available at discharge after the index surgery, such that resulting models would provide clinicians with a risk model for readmission at the time of discharge.

Materials and Methods

In order to mine patient records for relevant events, we used Progressive Insights which is a visualization toll that supports analysts searching for common patterns of events among large patient populations, as described in a previous paper.[10] Progressive Insights provides an integration of frequent sequence mining analytics with an exploratory visual interface with multiple coordinated views. We used Progressive Insights to identify the factors associated with 30-day readmission.

The frequent sequence mining analytics in this system are based on the SPAM algorithm.[11] The analytics are designed to identify the most frequent sequences occurring in a large patient population. Using Progressive Insights, users can specify the definition of frequent by controlling the minimum support threshold, which specifies the percentage of patients in which a pattern exists in to be considered frequent.

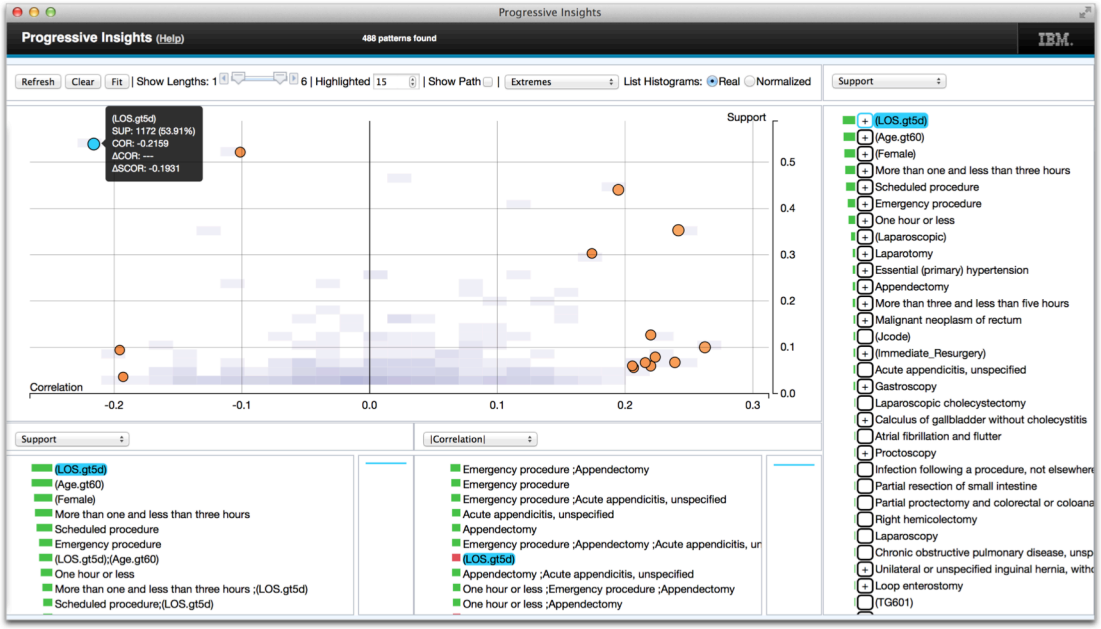


Figure 1: Screenshot of Progressive Insights. Progressive Insights supports analysts searching for common patterns of events among large patient populations. Progressive Insights provides an integration of frequent sequence mining analytics with an exploratory visual interface with multiple coordinated views, including two List Views (bottom), a Scatterplot View (center), and a Tree View (right). In the scatterplot, the horizontal axis is correlation, and the vertical axis is support. The 50 most frequent, correlated variables are plotted as orange circles. The presence of the other thousands of variables are plotted as a purple heatmap to reduce clutter, where the saturation of each cell corresponds to the number of variables in the region.

The results of the analytics are visualized in several interactive visualizations, including two List Views, a Scatterplot View, and a Tree View, as shown in Figure 1.The List View, shown at the bottom of Figure 1, is used in Progressive Insights to display the list of patterns detected by the SPAM algorithm sorted by one

of several user-selected ranking measures, such as support and correlation. Alongside each pattern is a histogram, where the bar length is proportional to the magnitude of the ranking feature. If the selected ranking is an outcome measure, such as Pearson correlation or odds ratio, the histogram's color indicates whether the pattern correlates with a positive (green) or negative (red) outcome.

The Scatterplot View, shown in Figure 1, displays patterns using user-selected metrics as the axes, so analysts can compare how metrics relate to each other within the data set. By default, the horizontal axis is correlation, and the vertical axis is support. Each pattern is plotted as an orange circle within the 2-dimensional scatter plot, and the size of the circle can be mapped to a third, user-selected metric (support, by default). However, depending on the minimum support threshold chosen by the user, hundreds or thousands of patterns may be discovered by the underlying analytics. In order to keep the scatterplot comprehensible, the system displays only the top *n* patterns as orange circles, ranked by user-selected metrics. Users can dynamically select various metrics and different values of *n* during their exploratory analysis. For all patterns that are not in the top-*n*, they are encoded by a heatmap in the background of the scatterplot. Each cell in the heatmap is colored using a log-scale color mapping, which indicates that the more purple a cell is, the more patterns that exist in that region of the scatterplot.

The Tree View, shown on the right of Figure 1, is designed to provide a hierarchical view for users to navigate the frequent patterns. The patterns are displayed in accordance with the tree hierarchy of prefix relationships inherent in the patterns. Concretely, this means that patterns that begin with an event type will be displayed as a root node in the tree. If users wish to explore patterns that begin with a certain event, users find that event in the tree and then expand it by clicking the '+' button next to the event. After expansion, users will see a list of events that frequently follow the selected event type. Users can apply this navigation technique to iteratively drill down into the patterns that are interesting. Like the List View, users can select different rankings to re-order the tree, and also feature a histogram representation of the selected ranking.

Although Progressive Insights features three distinct views for exploration, all of the views are coordinated. When users select a pattern in a view, the same pattern is selected in every view. This allows users to quickly leverage the distinct affordances of each view when relevant to their analysis.

Data source

We extracted the complete data from the EHR of a University hospital (University Hospital of North Norway; UNN) for all patients in the Department of Gastrointestinal Surgery. These data were managed in a database called QUAKE (Quality Control of Medical Performance by Unstructured EMR Data), containing both structured data and free text. Structured data was primarily diagnosis codes (International Classification of Diseases 10th edition; ICD-10) or procedure codes (Nomesco Classification of Surgical procedures; NCSP).[16]

A cohort of patients was selected, based on having an episode coded as a surgical event among those defined in Table 1. For all patients, their local ZIP codes were extracted from letters saved in the EHR, and the code closest in time to the time of surgery used. If the address was outside the region where UNN is the

Table 1: **Surgical procedures analyzed for 30-day readmission.** The top 10 types of surgeries included in the study, and number of patients in each group. Altogether 34 types were included, not all shown.

Type of surgery	Number of patients
Appendectomy	595
Open colon resections	566
Laparatomy	562
Laparoscopic cholecystectomy	446
Laparoscopic appendectomy	365
Open rectal procedures	347
Stoma formation	323
Laparoscopy	301
Inguinal and femoral hernia repair (open)	236
Open procedures small intestine	217

local hospital, the patients were excluded since readmission could happen to a different hospital. In total, 4307 patients were included in the final cohort. For all patients an index surgery was defined as the first episode containing any of the predefined surgical procedures. The most common procedures are summarized in Table 1. The index stay was defined as the period between admission and discharge that contained the index surgery. Readmission was defined as a non-outpatient hospitalization within 30 days after index surgery.

Variables

Variables that were included in the visual analytics were; diagnosis codes for a patient prior to and including the index stay, procedure codes after 60 prior to the index stay, demographic variables age (coded as greater than 60 years, close the median age for the population) and sex (coded as *female*). Additionally, factors believed to be associated with higher or lower readmission rates. These clinically defined factors were length of stay for the index stay (coded binary as longer or shorter than 5 days), intensive care treatment, previous abdominal surgery, previous heart surgery, laparoscopic surgery, and second surgery after the index surgery during the index stay[5]. Since we did not have data for other departments at the hospital such as the intensive care unit, we searched for the Norwegian word "Intensiven" as a proxy for whether patients were admitted to intensive care. For all variables a time stamp was included. Where the time point is not well defined (e.g., sex) the time of index surgery was used.

In Progressive Insights, all patterns were visualized and analyzed, including all that appeared in at least 1% of the cohort (minimum 43 patients). After visually assessing the result, interesting patterns were selected, for which the odds ratio (OR) and p-values were computed. Since we perform massive testing we correct for multiple testing by a Bonferroni correction, and an adjusted p-value computed.

The correlation in the visualization is the Pearson's correlation coefficient ρ. The statistic $\chi^2 = \rho N$ is χ^2 distributed with 1 degree of freedom, which allows estimation of the p-value for a particular pattern.

Results

A screenshot of the resulting visualization is shown in Figure 1. Interesting patterns that were identified along with the demographic and clinically defined variables are shown with their respective odds ratio and adjusted p-values in Table 2.

Table 2: Risk factors of readmission identified by visual analytics. Selected codes correlated negatively or positively with readmission ordered by absolute magnitude of correlation. The codes shown are those clinically specified (in italics) and selected codes appearing in the visualization as strongly correlating to the outcome and denoted as interesting by the clinicians. Patterns in bold are those with positive correlation, i.e., which reduce the chance of readmission.

Variable	N (%)	ρ	OR	p
Length of stay > 5 days	1847 (42.9)	-.256	0.301	< .0001
Malignant neoplasm of rectum	542 (6.9)	-.224	0.259	< .0001
Immediate resurgery	250 (5.8)	-.180	0.231	< .0001
Appendectomy	586 (13.6)	.140	3.48	< .0001
Operation time 3-5 hours	351 (8.15)	-.124	0.405	< .0001
Acute appendicitis,	293 (6.78)	.115	4.95	< .0001
Previous abdominal surgery	403 (9.36)	-.114	0.451	< .0001
Wound infection	197 (4.04)	-.114	0.345	< .0001
Laparoscopic procedure	1230 (28.6)	.113	1.87	< .0001
Proctoscopy	240 (5.2)	-.112	0.382	< .0001
Partial proctectomy with colorectal anastomosis	136 (3.16)	-.109	0.304	< .0001
Age > 60 years	2110 (49)	-.0957	0.345	< .0001
Laparoscopic appendectomy	365 (8.47)	.0943	2.70	< .0001
Reoperation for deep infection	45 (1.04)	-.0822	0.221	< .0001
Female	2148 (49.9)	.0418	1.21	0.12
Intensive care	177 (4.11)	-0.0332	0.700	0.65
Previous heart surgery	26 (0.6)	-.0168	0.636	1.0

There are a large number of points in the lower center of the diagram indicating codes or patterns that have low support and correlate little with the outcome. Patterns that rise up at the ends indicate clinically relevant variables connected to the outcome. In the interactive version, the user can click on any variable to obtain more information, or explore patterns of events happening together or in sequence.

Discussion

Progressive Insights provides an intuitive way to explore the patterns that correlate with an outcome in surgical practice by using data available in the EHR. By including as many factors as possible, the visualization allows both to confirm known patterns, and to explore new and unknown patterns. Importantly, this information is sometimes actionable in that unwanted outcomes can be reduced by changing routines based on institutional factors or by individual assessment of new patients based on uncovered risk factors.

Note that in our summary, each variable is evaluated separately to reflect how the visualization computes correlation. Adjusting for this by using, e.g., a multivariate model some effects may disappear if some of the variables are highly correlated with each other. The visualization should provide an initial insight into which variables to consider and not a complete statistical model. Hypothesis generation through visualization provides a starting point for pursuing any new insights and confirming them in subsequent rigorous analyses.

Surgery is a time intensive procedure where most information in the EHR arrives within a short time span, as contrasted to, e.g., chronic disease management that spans years or decades. This is a likely reason that most of the patterns uncovered in our visualization are simultaneous rather than sequential, which limits the utility of finding temporal patterns. Nevertheless, simultaneous patterns can convey useful information. By nature, concurrent patterns do not have a causal structure, and thus the sequence of events is not defined. In the current iteration of Progressive Insights, the resulting sequence is arbitrary, thus limiting the ability to distinguish concurrent events.

Clinical relevance

Identifying relationships between clinically relevant variables and clinical outcomes is important for clinicians and decision-makers. In our opinion, the use of novel visualization software tools may provide new insights in variables contributing to risk assessment models. The results showed that prolonged length of stay, stoma, high age, long operation time and second surgery was correlated with 30 days readmission (Table 2 and Figure 1). These variables are in accordance with numerous clinical surveys. Recently, Keller et al assessed factor associated with readmissions.[5] Preoperatively, they found that patients who were older, had more comorbidities, previous abdominal operations, and undergoing emergent procedures were at higher risk for readmission. In addition, patients who had open procedure, longer operative times, ICU stay, and longer LOS are more likely to be readmitted. Interestingly, these risk factors for 30-day readmission are in accordance with the factors visually identified to be correlated with readmissions, indicating that visualization tools provides clinically relevant information when assessing large datasets.[5,12,13]

There has been considerable debate regarding whether a hospital's 30-day hospital readmission rate is indeed a valid quality metric reflective of the care delivered at an individual institution. Recently Kiran and Delaney questioned the 30 day readmission rate as a adequate quality metric in colorectal cancer surgery, arguing that readmission within 30 days of a patient who has attained standardized discharge criteria may not be a valid indicator of poor quality of care.[8] In our analyses with Progressive Insights, we have assessed more than 2000 variables that might be associated with readmissions. In our opinion this provides a deeper insight in factors associated with 30-day readmissions after surgical procedures, and may have significant implications for both quality improvement initiatives and cost savings.[14] Furthermore, visual analytics can be an effective way for hospital decision makers to identify risk factors related to a certain clinical outcome.

Conclusion

In this paper, we have used a novel visual analytics tool to assess variables and their association with a highly clinically relevant and complex question; what characterizes patients that are readmitted within 30 days after a surgical procedure? The visualization provided a deeper insight in the most important factors associated 30-day readmission, and our results are in concordance with other surveys that uses traditional

statistical methods to identify these associations. The clinical problem is highly complex, with a wealth of risk factors impacting the outcome, and visual analytics provide an important additional component to perform in-depth analyses of this complex clinical question. This application's ability to visualize the association between two or more risk factors association with a clinical outcome is promising. In our opinion, visual analytics as a technique to identify unknown associations may play an important role in exploring and analyzing complex medical problems, and is a useful tool for decision-makers seeking actionable and preventable associations with unwanted outcomes.

Acknowledgements

Dr. Kristian Hindberg and Gisle Mjaatvedt are thanked for their effort in the data extraction process. SOS and KMA were partially funded through Tromsø Telemedicine Laboratory (TTL), which is a Center for Research-based Innovation in part funded by the Research Council of Norway grant no. 174934.

References

1. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA*. 2013;309(13):1351–2. doi:10.1001/jama.2013.393.
2. Cook KA, Thomas JJ, eds. *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society Press; 2005.
3. Tsai TC, Joynt KE, Orav EJ, Gawande AA, Jha AK. Variation in surgical-readmission rates and quality of hospital care. *N Engl J Med*. 2013;369:1134–1142. doi:10.1056/NEJMsa1303118.
4. Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: a systematic review. *JAMA*. 2011;306:1688–98. doi:10.1001/jama.2011.1515.
5. Keller DS, Swendseid B, Khorgami Z, et al. Predicting the unpredictable: comparing readmitted versus non-readmitted colorectal surgery patients. *Am J Surg*. 2014;207:346–51. doi:10.1016/j.amjsurg.2013.09.008.
6. Lawrence JK, Keller DS, Samia H, et al. Discharge within 24 to 72 hours of colorectal surgery is associated with low readmission rates when using enhanced recovery pathways. *J Am Coll Surg*. 2013;216:390–394. doi:10.1016/j.jamcollsurg.2012.12.014.
7. O'Brien DP, Senagore A, Merlino J, Brady K, Delaney C. Predictors and outcome of readmission after laparoscopic intestinal surgery. *World J Surg*. 2007;31:2430–2435. doi:10.1007/s00268-007-9345-3.
8. Kiran RP, Delaney CP, Senagore AJ, Steel M, Garafalo T, Fazio VW. Outcomes and prediction of hospital readmission after intestinal surgery. *J Am Coll Surg*. 2004;198:877–883. doi:10.1016/j.jamcollsurg.2004.01.036.
9. Wongsuphasawat K, Shneiderman B. Finding comparable temporal categorical records: A similarity measure with an interactive visualization. In: *IEEE Symposium on Visual Analytics Science and Technology*. Atlantic City, New Jersey, USA: IEEE; 2009:27–34. doi:10.1109/VAST.2009.5332595.
10. Stolper CD, Perer A, Gotz D. Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics. *IEEE Trans Vis Comput Graph*. 2014:Accepted.
11. Ayres J, Flannick J, Gehrke J, Yiu T. Sequential Pattern mining using a bitmap representation. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*.; 2002:429. doi:10.1145/775107.775109.
12. Kassin MT, Owen RM, Perez SD, et al. Risk factors for 30-day hospital readmission among general surgery patients. *J Am Coll Surg*. 2012;215:322–330. doi:10.1016/j.jamcollsurg.2012.05.024.
13. Sweeney JF. Postoperative Complications and Hospital Readmissions in Surgical Patients. *Ann Surg*. 2013;258(1):19–20. doi:10.1097/SLA.0b013e318297a37e.
14. Lawson EH, Hall BL, Louie R, et al. Association between occurrence of a postoperative complication and readmission: implications for quality improvement and cost savings. *Ann Surg*. 2013;258:10–18. doi:10.1097/SLA.0b013e31828e3ac3.

Visual Hypothesis & Correlation Discovery for Precision Medicine

R. Wolfgang Rumpf, PhD[1], Jenn Gonya, PhD[1,] William C. Ray, PhD[1]
[1]The Research Institute at Nationwide Children's Hospital, Columbus, OH

Abstract

Hospitals collect a tremendous amount of information on patients, including basic metrics, vitals, and outcome scores. This information is effectively useless if it cannot be used to identify trends in patient responses and outcomes based on condition and treatment trajectories, and thereby provide a way to improve treatment. Unfortunately, because the data is highly heterogeneous and frequently contains dense networks of dependencies between both the measured variables and unmeasured latent variables, this wealth of data is usually cherry-picked for only a handful of the overwhelmingly obvious features. This approach significantly impedes understanding what features are relevant to an individual in a Precision Medicine context. One approach to improving this situation is to enable the simultaneous visualization of all marginal and joint distributions of the measurements as seen in the population, and enable projection of an individual's data into this statistical space, so that the optimal approach for that patient based on their individual fit to the population trends can be determined. We have developed a visual tool, StickWRLD, loosely based on the idea of categorical parallel coordinates, which can depict the full marginal, and pairwise, tertiary, or quaternary joint distributions of numerous heterogeneous measurements across a population. Originally developed for understanding the effects of networks of mutations in specific proteins based on population data, here we have adapted the approach to visually analyzing a 57-variable NICU dataset detailing premature infants' responses to various forms of skin-to-skin (kangaroo) treatment.

Introduction

The information that hospitals collect on their patients typically consists of a vast number of qualitative and quantitative "vital signs" and test measurements as well as outcome scores. The resulting data sets contain numerous variables and possibly conditional dependencies between variables. While pure machine-learning approaches can yield accurate predictors that suggest optimal treatment plans for an individual, these approaches are effectively "black boxes" that do not facilitate understanding of the prediction, or enable expert intuition regarding whether an individual's unique characteristics have been appropriately weighted in the decision making process. Only through an interactive visual exploration of the statistical structure that an understanding of the patterns and relationships can be revealed[1,2].

While many attempts at interactively visualizing patient records (including Electronic Health Records, or EHRs) focus on the concise display of single-patient statistics aimed at the healthcare provider[3,4], other systems are research-centric, allowing researchers to examine multi-patient data[5]. By screening such multi-patient datasets, researchers can discover which factors have an impact on outcome measures, leading to better, directed interventions – provided they are able to explore the complete dataspace of possible interactions. With the large number of possible factors to consider, a means to rapidly screen the many possible variables is required so that further efforts can be focused on refining specifically how those variables can impact patient outcome.

StickWRLD is a visual analytic tool originally designed for visually exploring functionally-required networks of residue dependencies derived from protein family sequence data (multiple sequence alignments), and how the properties of these relationships play out in individual proteins, differentially affecting how they respond to mutations[6]. StickWRLD has also been used to explore networks of SNPs and gene expression levels in an eQTL dataset[7]. Here we describe the application of StickWRLD to a new data type: multi-patient clinical dataset for the purpose of understanding clinical variables that are similarly interdependent and affect outcome scores in a positive or negative fashion.

Visualization Design

StickWRLD is a dynamic visualization tool that allows users to "browse" through statistical features

(primarily marginal and joint distribution features, though other measures of relationship can be encoded) of the population and individuals in that population, graphically displaying "interesting" relationships between variables in an interactive 3D interface. StickWRLD's default calculation assigns a magnitude of "interest" based on the difference between the expected joint distribution for variables, and the observed joint distribution (effectively the residual), though any calculation of relatedness can be injected into the visualization. On top of these statistics, StickWRLD layers an interface in which the user can interactively explore the consequences of different combinations of thresholds and heuristics for reducing the quantity of interesting features displayed. Through this interaction StickWRLD essentially becomes a hypothesis generator, enabling users to weed through thousands of possible patterns of correlations and graphically filter the signal from the noise, discovering patterns of interest in population data, based both on statistical properties as well as the domain expertise of the user, and to examine how an individual fits the population patterns. StickWRLD and the human brain's pattern-recognition ability effectively become a visual analytical engine that can be used to develop specific hypotheses to test, or to identify specific strategies for individualized treatment. Because StickWRLD focuses attention on unexpected features of joint distributions, its visual approach readily highlights non-linear dependencies (for example, threshold effects that are often overlooked in traditional regression analyses.

StickWRLD displays each variable in a dataset as a column, with columns arranged in a ring to form a cylinder. This cylindrical representation allows StickWRLD to display correlations between any possible combinations of variables. Within each column, the marginal distribution of values observed for that variable is displayed. Each possible value is represented as a sphere, with the size of the sphere indicating the frequency of occurrence of that particular value within the dataset (e.g. over-represented values are seen as large spheres, whereas under-represented values are seen as small spheres). Correlations between variables are displayed as a cylinder connecting the appropriate spheres, showing the user not only which variable is positively (solid line) or negatively (dashed line) influenced by which other variable(s), but also displaying which values or categories within each variable, specifically, are linked. The strength of the relationship is indicated by the thickness of the line. By varying statistical and magnitude thresholds, users can manually drive the display to visualize more – or less – significant relationships. The entire interface is a user-directed 3D representation, allowing the user to zoom in and out as well as to rotate and pan the display to explore and home in on specific relationships.

This representation is essentially an interactive 3D projection of traditional parallel coordinates visualization. Parallel coordinates are frequently used to visualize dependencies within high dimensional data, since every point in the high-dimensional space can be unambiguously represented in the parallel coordinates plot by a distinct polyline. While more traditionally used for continuous-value data, parallel coordinates have also been applied to categorical data[8,9] such as the data presented here. Traditional parallel coordinates representations however suffer from two distinct failings that make them difficult to meaningfully apply in an exploratory/analytics context such as Precision Medicine. The first is that, barring axis duplication which invokes additional issues, traditional parallel coordinates methods enforce a *mandatory* sequential ordering on the data, and can only represent dependencies between a variable and its immediately preceding or following neighbors. The second is that for a variable and its neighbors, the display is effectively static - *all* dependencies, regardless of interest, are always shown. StickWRLD overcomes these limitations both through its interactive analytics paradigm, and through its layout[10].

By wrapping parallel coordinate axes into a cylinder, StickWRLD violates the conventional wisdom that good visualizations should constrain themselves to two dimensions (to avoid the subconscious perception that foreground objects are more relevant), and introduces issues where edges may be occluded dependent on the viewpoint. We justify the use of 3D in two ways: First, since StickWRLD is an interactive visualization, there is no privileged perspective to mislead the user, and second, by moving the parallel coordinates from a plane into a volume, StickWRLD allows for the simultaneous comparison of all possible relationships – reducing the time required to analyze the relationships to minutes, whereas a more traditional multiple-2D graph approach would take significantly longer[10]. StickWRLD, and other extensions of parallel coordinates into a 3D space[11-13], all take different approaches to solving the problem of edge occlusion. StickWRLD's solution is to display only edges that meet or exceed a user-defined

residual, or *observed – expected* threshold, which is by default set arbitrarily high. This minimizes potentially occlusive "clutter" while drawing out the strongest dependencies for consideration.

While this layout and approach is somewhat controversial in the Visualization world, and there are clear scaling issues that make it most appropriate for data sets of only up to several hundred variables (summarization, clustering and variable-hiding features not discussed here extend this range to several thousand variables), we have demonstrated that StickWRLD's approach provides a uniquely powerful tool for developing otherwise unattainable statistical models of densely connected interacting feature networks in protein sequence data, an application domain that is strikingly similar to that proposed here[10].

There is significant ongoing discussion in the field of visual analytics pertaining to optimal methodologies for visualizing datasets with very large numbers of nodes, edges, and edge-crossings[14-17]. We address this general problem in two ways: StickWRLD's interactive 3D cylindrical visualization allows the user to enter and navigate through the data space, while dynamically modifying the number of displayed edges by varying the p and residual thresholds. This allows the user to examine the data space at any desired level, focusing in on relationships of interest as defined by either their statistical significance, or their significance as defined by the domain expertise of the user. More importantly, however, StickWRLD takes the approach that, in some cases, the noise and clutter are in fact desirable – StickWRLD takes advantage of the human brain's pattern-recognition ability to see patterns as they emerge during the user's dynamic modification of the statistical model.

Case Study

Extremely premature infants are one of the highest risk categories for developmental impairment due to the adversity that comes from missing the third trimester in utero[18]. Much interest has been generated within the neonatology community for evaluating the developmental outcomes of such infants and understanding how these outcomes are affected by practices within the Neonatal Intensive Care Unit (NICU). The challenge is how to investigate the effects of a multitude of variables that neonates encounter during their hospitalization. To apply StickWRLD to this problem, NICU data for 38 separate measures (including standard measures such as gestational age, birth weight, gender, time spent on a ventilator (IPPV), days until mouth feeding supplants intubated feeding, etc.) was collected for 57 premature infants over a one-year period. For each measure, the values were binned appropriately, using multiple schemes when there was no single obvious categorical way of binning the data. All bins were represented with alphabetic designations to facilitate entry into StickWRLD. The complete datasets (original as well as binning variations) as well as the StickWRLD python scripts and manual, which contains instructions for preparing StickWRLD format datasets, are available for download from *http://www.stickwrld.org/VAHC-2014/*

Results

Initial binning dataset

The initial StickWRLD view of the dataset validated the approach, displaying several expected correlations. In Figure 1A, the Length of Stay (LOS, right forefront) is strongly correlated with the number of days until full mouth feeding (PO DOL, left forefront). This relationship is unsurprising, as premature infants are typically not discharged from the NICU until they can feed exclusively by mouth. Figure 1B shows a different perspective, with the inter-correlation of the 6-month Bayley outcome test scores (cognitive, communication, and motor) in the foreground – an infant with an average (or above average) score in any of these categories tended to have an average (or above average) score in all categories.

StickWRLD also displayed a clear linkage between the presence of intraventricular hemorrhage (IVH) and Grade 1 IVH. This relationship demonstrates that caution must be taken when binning clinical data for visualization in StickWRLD – while the relationship is valid, it simply shows that the majority of IVH cases are Grade 1. Rather than encoding a binary (yes/no) state for the presence or absence of a condition in one variable, and the degree of the condition in a second variable, one variable for the degree with

"none" as one possible value would reduce the complexity of the analysis, resulting in fewer "obvious but uninteresting" correlations such as the one between the *presence* of IVH and the *severity*.

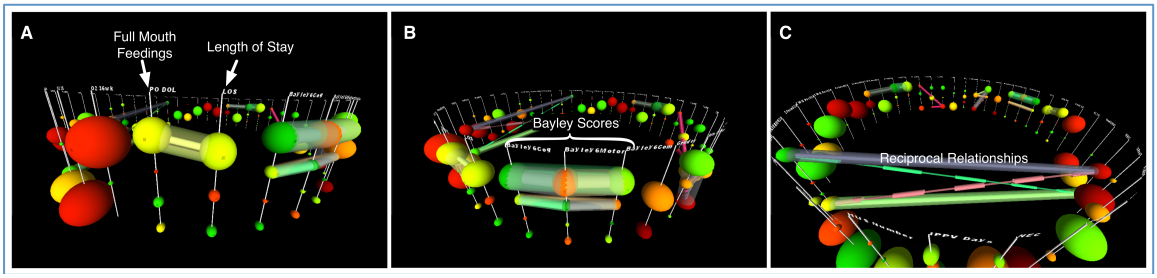


Figure 1. StickWRLD view of the initial dataset.

One of the values of a rapid dynamic visualization such as StickWRLD is the ability to browse the data looking for relationships of interest. Not displayed are several relationships discovered when the thresholds for *p* and *r* were tuned to even lower levels – for example, at *p=0.06* and *r=0.06*, a correlation was seen between low birth weight and low Bayley scores (which, while not amenable to NICU intervention nor statistically significant, suggests that at-risk mothers should be encouraged to take steps to ensure as high a birthweight as possible for their infants). Another correlation seen at these settings linked treatment of PDA (but not the disease itself) to longer NICU stays. This sort of unexpected discovery is only possible because of StickWRLD's simultaneous computation and display of correlations between all possible nodes.

Binning variation #2

To see the impact different binning schemes can have on the visualization, we re-binned the dataset with fewer, broader bins for several of the non-outcome variables. Visualization of the newly derived dataset revealed a new correlation – that of the onset of time of kangaroo care (KCONSET) to average Bayley scores. Figure 2 shows a top-down view of the visualization (Fig 2 Panel A, with KCONSET at the bottom and Bayley Scores are at the top) as well as an orthogonal view (Fig 2 Panel B, with KCONSET in the foreground and Bayley scores in the background). StickWRLD shows a strong correlation between very early (0 to 5 days after admission) onset of Kangaroo Care and average (71-100) Bayley scores.

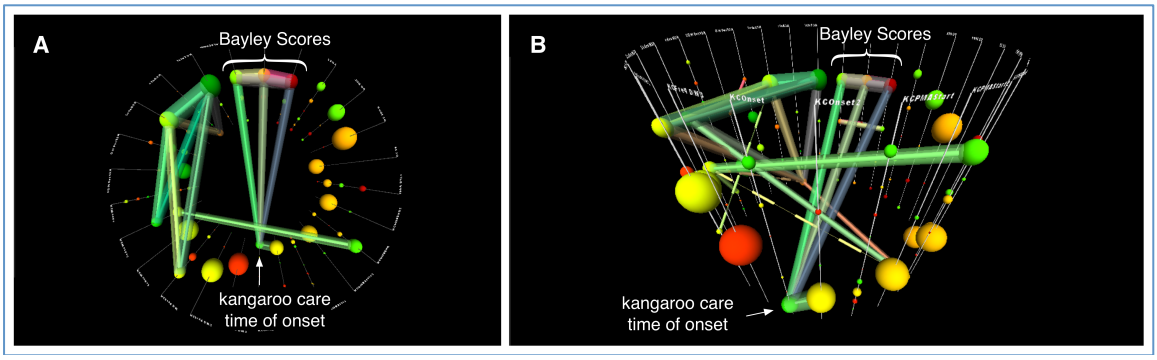


Figure 2. StickWRLD visualization of Binning Variation #2

Figure 2 also reveals an additional "uninteresting/expected" correlation – total length of kangaroo care in days is correlated to the time of onset of kangaroo care. While real, this is inherently obvious and does not suggest any relationship between kangaroo care and outcome measures. This again supports the need for domain expertise in both binning of datasets as well as subsequent visual analysis.

Binning variation #3

To attempt to put more granularity into the correlation between kangaroo care onset and Bayley scores seen in Binning variation #2, a third binning variation was generated with Bayley scores broken into a larger

number of smaller bins. When this variation dataset was visualized, the correlation between kangaroo care onset and Bayley scores vanished, indicating that the correlation was to the larger range of values rather than to a narrower range. A new correlation was seen between 0-30 days on the ventilator, or IPPV (Figure 3, background), and above-average (101-110) Bayley Communication scores (Fig 3, foreground).

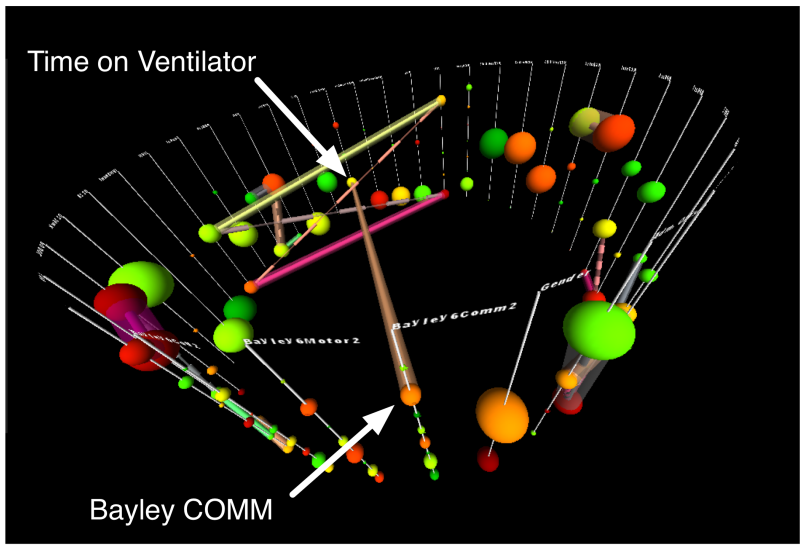


Figure 3. StickWRLD visualization of Binning Variation #3 dataset.

Using StickWRLD as an hypothesis engine, we can now generate a testable hypothesis from the discovered correlations. From the original, unbinned data we were able to isolate an individual with a low number of kangaroo care days whose IPPV and Bayley Communication score were low (contrary to the correlation seen in Figure 3). A caregiver in the NICU with access to this analysis might have been able intervene by increasing the number of kangaroo care days for this individual, testing whether this resulted in an increase in their Bayley Communications score. While only possible as a thought experiment for this particular infant, who has long since left the NICU, it demonstrates how StickWRLD can be used to sift through a clinical dataset looking for correlations between possible interventions and outcome measures.

Conclusions

StickWRLD, originally designed to visualize residue coinheritance/correlation in bioinformatics sequence datasets, can be applied to clinical health data. StickWRLD calculates the strength of correlation for all possible combinations of variables in a dataset and then displays those exceeding the user-defined threshold settings – and dynamically allows the user to change those thresholds, updating the display in real-time. Because clinical data tends to be comprised of continuous and/or quantitative variables, rather than the finite set of discrete values possible in bioinformatics sequence data, care must be taken when binning the data into StickWRLD format. Both construction and subsequent analysis of a clinical dataset benefit from the application of domain expertise.

References

1. Gershon N, Eick SG. Visualization's new tack: Making sense of information. Spectrum, IEEE 1995;32:38-40, 2, 4-7, 55-6.

2. Shahar Y, Goren-Bar D, Boaz D, Tahan G. Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. Artif Intell Med 2006;38:115-35.

3. Plaisant C Milash B, Rose A, Widoff S, Shneiderman B. LifeLines: Visualizing Personal Histories. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1996:221-227.

4. Plaisant C, Mushlin R, Snyder A, Li J, Heller D, Shneiderman B. LifeLines: using visualization to enhance navigation and analysis of patient records. Proc. AMIA Symp. 1998:76-80.

5. Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. Artif. Intell. Med. May 2010;49(1):11-31.

6. Ray, WC. MAVL and StickWRLD: visually exploring relationships in nucleic acid sequence. NAR 2004;32(Web Server issue):W5-63.

7. Ray, WC. StickWRLD: Interactive Visualization of massive parallel contingency data for Personalized Analysis to facilitate Precision Medicine. 2013 AMIA Workshop on Visual Analytics in Health Care.

8. Rosario GE, Rundensteiner EA, Brown DC, Ward MO, Huang S: Mapping nominal values to numbers for effective visualization. Inform Visual 2004, 3(2):80–95.

9. Bendix F, Kosara R, Hauser H: Parallel sets: visual analysis of categorical data. In Information Visualization, 2005. INFOVIS 2005. IEEE Symposium On. New York, NY: IEEE Press; 2005:133–140.

10. Ray, WC, Wolock, SL, Callahan, NW, et. al. Addressing the unmet need for visualizing conditional random fields in biological data. BMC Bioinformatics 2014;15:202.

11. Fanea E, Carpendale S, Isenberg T: An interactive 3d integration of parallel coordinates and star glyphs. In Information Visualization, 2005. INFOVIS 2005. IEEE Symposium On. New York, NY: IEEE Press; 2005:149–156.

12. Johansson J, Ljung P, Jern M, Cooper M: Revealing structure in visualizations of dense 2d and 3d parallel coordinates. In Inf Vis. Thousand Oaks, CA: SAGE Publications; 2006.

13. Kerren A, Jusufi I: 3d kiviat diagrams for the interactive analysis of software metric trends. In Proceedings of the 5th International Symposium on Software Visualization. SOFTVIS '10. New York: ACM; 2010:203–204. [http://doi.acm.org/10.1145/1879211.1879241]

14. Rosenholtz, R, Li, Y, Mansfield, J, Jin, Z. Feature congestion: a measure of display clutter. In: Proc. SIGCHI Conference on Human Factors in Computing System 2005;761-770.

15. Rusu, A, Fabian, AJ, Jianu, R, Rusu, A. Using the gestalt principle of closure to alleviate the edge crossing problem in graph drawings. In: Proc. International Conference on Information Visualization 2011;488-493.

16. Burch, M, Vehlow, C, Konevtsova, N, Weiskopf, D. Evaluating partially drawn links for directed graph edges. In: Proc. Of the 19th international conference on Graph Drawing 2011;pp 226-237.

17. Von Landesberger, T, Kuijper, A, Schreck, T et al. Visual analysis of large graphs: state-of-the-art and future research challenges. Computer Graphics Forum 2011;30(6):1719-1749.

18. Als H, Duffy FH, McAnulty GB, Rivkin MJ, et al. Early experience alters brain function and structure. Pediatrics 2004;113:846-57.

WorkflowExplorer: Visual Exploration and Identification of Common Multitasking Patterns in Emergency Department Workflow

Allan Fong, MS[1,2], Kevin Maloy, MD[2,3], Raj Ratwani, PHD[1,2,4]
[1]National Center for Human Factors in Healthcare, Washington DC; [2]MedStar Institute for Innovation, Washington DC; [3]Department of Emergency Medicine, Washington Hospital Center, Washington DC; [4]Georgetown University School of Medicine, Washington, DC

Abstract

Understanding workflow in healthcare settings is important to the safety and timely care of patients. However, discovering patterns in complex workflow data, such as those associated with the emergency department, can be challenging. Task stacking, multitasking, interruptions, and task weaving can obscure trends in common workflow patterns, yet understanding these processes is critical to understanding workflow. In this paper, we describe WorkflowExplorer, a dense data visualization designed to help users explore and find common transition sequence patterns by identifying and highlighting common sequences that meet certain user requirements. We present the interactions and capabilities of the visualization as well as the underlying probabilitistic model. We briefly demonstrate WorkflowExplorer's utility at identifying common patterns in the workflow of emergency medicine physicians across five observations.

1. Introduction

Emergency medicine (EM) physicians deliver unscheduled, high acuity care to multiple patients simultaneously. They have little to no control over incoming workflow, volume, or acuity. Furthermore, they often make decisions with incomplete data. As a result, developing a comprehensive understanding of EM workflow is challenging, yet a deep understanding of workflow is critical to the safe, efficient, and effective delivery of care to the patients[1,2]. In this paper, we present WorkflowExplorer, a visualization tool we developed to identify common transition sequence patterns in EM physician workflow. We discuss the tool's development, interactions, and how it can be utilized to analyze EM physician workflow data.

2. Background

2.1 Workflow in the ED

Understanding physician workflow, which includes the work tasks performed and the tools and technologies used to accomplish those tasks, is critical to improving safety, quality and efficiency. In complex settings like the emergency department (ED), workflows are particularly difficult to understand. ED clinicians manage multiple patients with rapidly changing conditions and must be ready to adapt to accommodate new patients that arrive unexpectedly[3]. Given the conditions of the ED, physicians are often forced to perform multiple tasks simultaneously and must strategically interleave tasks and handle interruptions to ensure that appropriate care is delivered in a timely fashion. The literature examining performance under conditions of multitasking and interruptions has demonstrated how error prone these processes can be[4-6].

Most studies examining ED workflow have focused on quantifying the types of tasks performed and the length of time spent on different tasks[7-9]. These first order descriptives provide a good and necessary high-level characterization of workflow. However, a deep understanding of the complexities and nuiances of work processes can be difficult to achieve from such descriptive data. For example, it is difficult to glean information on the specific temporal patterns of different tasks from aggregated descriptive data and it is unclear as to which tasks may be performed in parallel or serially. Advanced data visualization techniques have the potential to provide these more complex details in a format that is easy to interpret.

2.2 Understanding Workflow through Visual Analytics

Emergency medicine (EM) physician workflow is an example of complex multi-dimensional temporal data. Properly designed information visualization systems can greatly assist in the exploration and discovery of patterns and relationships[10-12]. Systems that leverage both the human's ability to detect trends and the computer's ability to rapidly process data can help users discover, test, and evaluate meaningful and significant patterns[13]. There have

been many systems and concepts developed to assist in the visualization and understanding of complex temporal data, such as LifeFlow, EventFlow, VizTree, ThemeRiver, and others[14-18]. While these visualizations offer novel and useful ways to explore temporal data, it is challenging to find an integrated tool that both captures the intricate relationships of multiple interweaving tasks and identifies common transition patterns for the user. Tools to help users discover temporal patterns and trends can be greatly enhanced by incorporating statistical and machine learning techniques.

In this paper, we discuss WorkflowExplorer, a visualization developed utilizing a state transition probability model to automatically highlight common transition sequences based on user defined parameters. This visualization tool was developed based on challenges when exploring complex workflow data to identify meaningful patterns. There is a plethora of algorithms and approaches already developed for sequence pattern identification that can aid in this process, such as Hidden Markov Models (HMM), Symbolic Aggregate approXimation (SAX), Bowtie, prefix and suffix trees[19-23]. VizTree and PairFinder are examples of visualizations that incorporate more advanced algorithms or statistics to assist users in the overall exploration and discovery process[16,24]. In this paper, we expand on this work and developed a tool that makes the identification of common transition sequences in workflow data easier and more intuitive. We discuss how we incorporated conditional state transition probabilities into our visualization and how we developed an user interface to help facilitate the exploration and discovery of patterns and trends in EM physician workflow.

3. Approach

In this section, we discuss the data used, our state transition probability model, and the features and interactions of WorkflowExplorer, as seen in Figure 1.

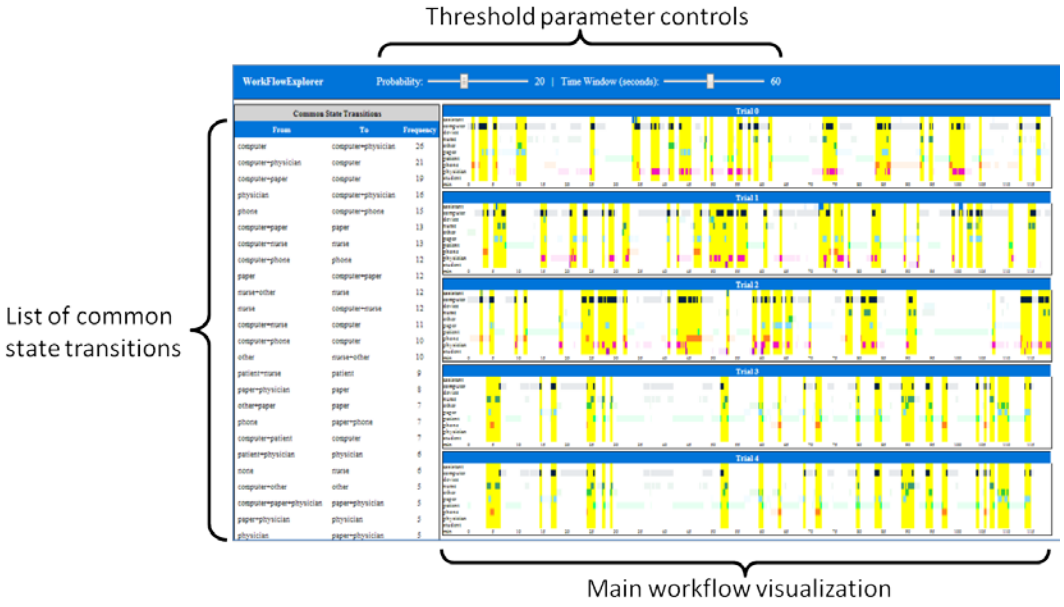


Figure 1: WorkFlowExplorer helps users explore and find common state transition sequences

3.1 Workflow Data

We tracked the workflow of emergency medicine (EM) physicians at various hospital sites using a tablet running a web application designed to collect emergency room workflow data[25]. We captured when participating physicians were 1) working on the computer [Computer], 2) reading or documenting something on paper [Paper], 3) directly taking care of a patient [Patient], 4) talking on the phone [Phone], 5) using another device [Device], 6) talking to a nurse [Nurse], 7) talking to a student [Student], 8) talking to an assistant or technician [Assist/Tech], 9) talking to another physician [Physician], 10) performing any other task [Other], or some combination of the above. Each unique combination of task(s) is considered a different state. For example, talking to a nurse, [Nurse], versus talking to a nurse while documenting information on a computer, [Nurse + Computer], are two different states. This allowed us to track when physicians switched tasks as well as when they were multitasking. This continuous data was discretized into ten second increments[22]. This provided some data smoothing and made categorizing multitasking

intervals easier. The example in Figure 2 illustrates the resulting data structure (right) after the discretizing and smoothing process. We used this discretized data to develop our state transition probability model and visualization.

Start time (s)	End time (s)	Event
2	19	Computer
11	51	Paper
48	68	Computer

Time (s)	Computer	Paper	State
0	1		1
10	1	1	3
20		1	2
30		1	2
40	1	1	3
50	1		1
60	1		1

Figure 2: Continuous data (left) discretized into discrete states (right)

3.2 Defining the State Transition Probabilistic Model

The state transition probabilities associated with each unique state change lay the foundation for our visualization. We define the transition probability of being in a state, s_i (i=1, 2, 3,...N where N is the total number of states), as simply the conditional probability of arriving at the current state (s_i) given the previous state (s_j) or:

$$p(s_i|s_j) = \frac{p(s_j|s_i)p(s_i)}{p(s_j)}$$

where $p(s_i)$ and $p(s_j)$ are the transition probabilities of being in states s_i and s_j, respectively, given all the other states. As an illustration, in Figure 2, the probability at time 10s would be the conditional probability the participating physician would switch to a [Computer + Paper] multitasking state after working on just the [Computer]. A transition probability was assigned for each time interval associated with a state change, excluding transitions to the starting state and time intervals that did not change from the previous state.

3.3 Guided Discovery of Transition Subsequences

Our visualization helps users explore and discover patterns in workflow data by highlighting and summarizing common transition sequences based on user defined threshold values, Figure 3.

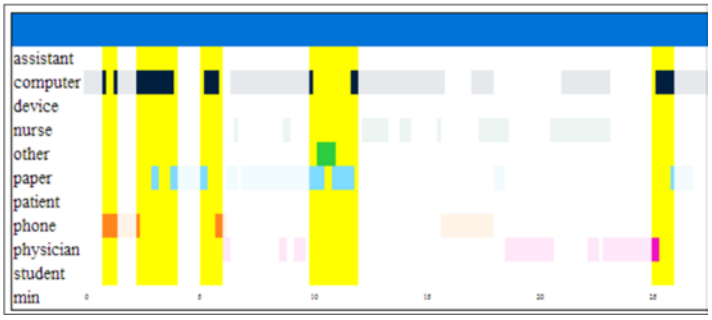


Figure 3: Common sequence patterns in this sample workflow data are highlighted in yellow

Transition probability threshold

WorkflowExplorer lets users adjust the state transition probability threshold (f_p) based on their experience, interest, etc., Figure 4 (left). Setting this parameter at a low value, f_p = 20 (or 0.2), will consider any transition probabilities greater than 20 (or 0.2) as important to the user. A low f_p will tend to highlight more common transition sequence than higher f_p values.

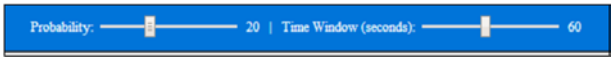


Figure 4: User controls to adjust the state transition probability threshold (left) and the time window threshold (right)

Time window threshold

The time in which transitions occur, or time window of transitions, is another important measure to consider. Although a sequence of transitions could be common, such as [NURSE + PAPER] to [PAPER] to [PAPER +

COMPUTER], if the times between the transitions were long or greatly varied, the transition sequence might be less meaningful. Taking this into account, our visualization lets users adjust a time window threshold (f_t) which defines the maximum distance between two important transitions, as defined by f_p, required to consider the interval an interesting sequence. For example, two state transitions greater than f_p 15 seconds apart, would not be highlighted as a common transition sequence if f_t = 10s, but would be highlighted if f_t = 20s. This interaction lets users explore how time intervals affect the identification of patterns.

Identifying common transition sequence patterns

WorkflowExplorer identifies and highlights common transition sequence patterns in workflow data that meets the parameter thresholds defined by the user. This is accomplished simply by identifying intervals, f_t or less, that contain two state transitions of probabilities, f_p or greater. However, if there are clusters of relevant state transitions, the resulting sequence pattern could be longer than f_t. For example, in Figure 5, the number of unique patterns depends on the time window threshold, f_t. If f_t = 20s, two separate transition sequence patterns would be highlighted. However, if f_t = 50s, only one longer transition sequence would be highlighted. As a result, the length and count of transition sequence patterns identified greatly depends on both the f_p and f_t thresholds.

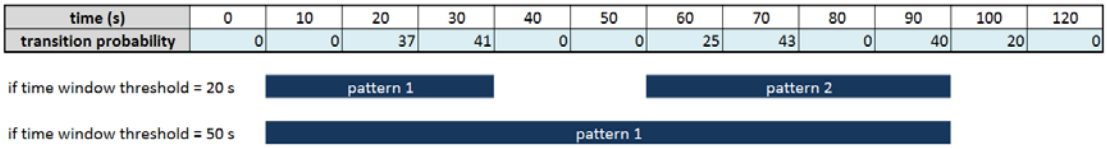


Figure 5: The length and count of identified patterns will depend on the time window threshold

Furthermore, the common state transitions representative of the highlighted patterns are listed in the visualization. The summary table in the left panel of the visualization provides an overview of the top transitions given the user parameters, Figure 1. The common transitions are displayed as text to conserve space and make the trends more readable. This can help users both generate hypothesis as well as find and explore different trends and patterns.

4. Interactive Exploration and Discovery of Patterns

With WorkflowExplorer, users can easily and quickly identify workflow patterns and explore the sensitivity of these patterns by adjusting the transition probability and time window thresholds. The visualization examples in Figure 6 show the highlighted EM physician workflow patterns using different parameter settings.

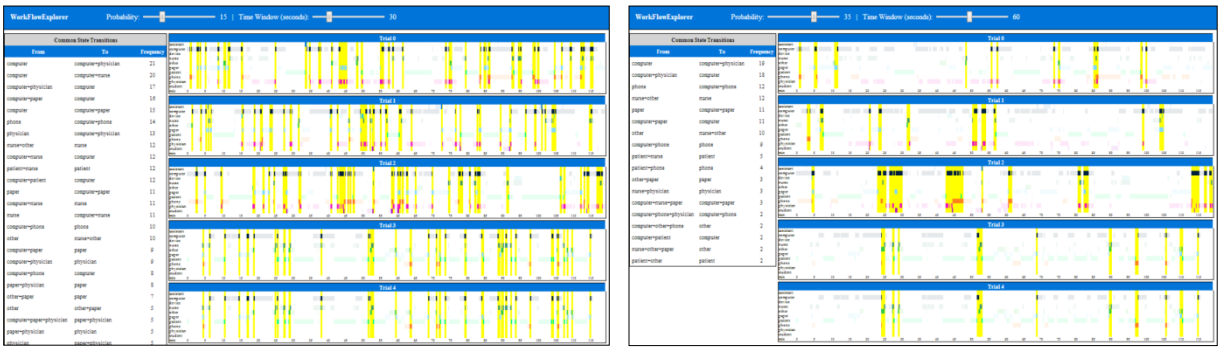


Figure 6: Common transition sequence patterns results with different threshold parameters

By setting f_p to 15 and f_t to 30s, Figure 6 (left), WorkflowExplorer identifies a number of patterns where the physician is switching between one and multiple tasks. This clearly highlights the highly dynamic nature of EM physician workflow. We also notice the prevalence of multitasking and task switching centered around the computer. As hospitals adopt electronic medical systems, it is expected that computers are going to be used more in the patient care process. To further investigate the more probable and longer transition sequences, we increase f_p and f_t, Figure 6 (right). Some initial insights we immediately observe are the differences between the five observation sessions. The fourth and fifth observation session have clearly different workflow patterns than the other three. It is, for example, interesting to observe that the common patterns in the fourth and fifth observations involved nurses much more than the other three. This visualization can quickly highlight common transition sequence patterns for users to explore and investigate further.

5. Initial Feedback and General Discussion

We used our tool to visualize transition trends and patterns in emergency medicine (EM) physician workflow. We solicited initial feedback regarding WorkflowExplorer from various researchers interested in understanding temporal patterns in workflow data. Users commented on how the highlighted common transition sequence patterns provide an useful starting point for data exploration. Users enjoyed the simplistic threshold controllers and how the results were updated in real-time. We also received various feedback and suggestions that will greatly improve the interactions and visualization of the tool. These suggestions include additional user controlled parameters, such as the minimum number of transitions required for any given sequence. The current default system requires at least two transitions per pattern. In addition, it would be helpful to provide more descriptions of the common transition sequence patterns in the summary window.

Furthermore, in WorkflowExplorer, the most probable state changes are transitions to and from multitasking states rather than from one task directly to another task. This initial insight highlights the rapid task weaving and multitasking nature of EM physician workflow. This visualization also lays the foundation for many additional unique interactions and analyses. It would be useful, for example, to visualize when external interruptions occur and highlight common workflow behaviors leading up to multitasking events. Understanding what leads up to multitasking events or how physicians respond to interruptions can have important implications on patient safety and patient care.

6. Conclusion

Understanding workflow is important to the evaluation and improvement of patient care and patient safety, particularly in complex highly dynamic environments, such as the emergency department (ED). However, studying workflow patterns in the ED is very difficult because of the temporal and multitasking nature of the work. We built WorkflowExplorer to help users better understand emergency medicine (EM) physician workflow through a guided exploration of the data. Our visualization helps the user explore and identify workflow patterns base on the user's experience, domain knowledge, and thresholds. We demonstrated how this visualization could identify different workflow patterns based on a user's threshold settings. We provide initial user feedback of the visualization as well as suggestions for future improvements.

7. Acknowledgements

We would like to thank the physician participants. This work was supported by AHRQ grant # 1 R03 HS022362-01 to Raj Ratwani.

References

1. Fagerhaugh SY. Social organization of medical work. Transaction Publishers. 1997.
2. Lenz R, Reichert M. IT support for healthcare processes – premises, challenges, perspectives. Data and Knowledge Engineering. 2007; 61(1): 39-58.
3. Ratwani R, Hettinger AZ, Fairbanks R. Human Factors in Emergency Care. In Emergency Care and The Public's Health, First Edition. Wiley & Sons. 2014; (in press).
4. Drews FA. The frequency and impact of task interruptions in the ICU. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2007; 51(11): 683-686.
5. Trafton JG, Altmann EM, Brock DP, Mintz FE. Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. International Journal of Human-Computer Studies. 2003; 58(5): 583-603.
6. Monk CA, Trafton JG, Boehm-Davis DA. The effect of interruption duration and demand on resuming suspended goals. Journal of Experimental Psychology: Applied. 2008; 14(4): 299.
7. Chisholm CD, Collison EK, Nelson DR, Cordell WH. Emergency Department Workplace Interruptions Are Emergency Physicians "Interrupt-driven" and "Multitasking"? Academic Emergency Medicine. 2000; 7(11): 1239-1243.
8. Westbrook JI, Coiera E, Dunsmuir WT, Brown BM, Kelk N, Paoloni R, Tran C. The impact of interruptions on clinical task completion. Quality and Safety in Health Care. 2010; 19(4): 284-289.
9. Berg LM, Källberg AS, Göransson KE, Östergren J, Florin J, Ehrenberg A. Interruptions in emergency department work: an observational and interview study. BMJ Quality & Safety. 2013; 22(8); 656.
10. Keim DA, Andrienko G, Fekete J, Gorg C, Kohlhammer J, Melancon G. Visual Analytics : Definition , Process, and Challenges. Springer Berlin Heidelberg. 2008; 154-175.

11. Keim DA, Mansmann F, Thomas J. Visual Analytics: How Much Visualization and How Much Analytics? ACM SIGKDD Explorations Newsletter. 2010; 11(2): 5-8.
12. Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. In Visual Languages. 1996; 336-343.
13. Wong PC. Guest Editors' Introduction: Visual Data Mining. IEEE Computer Graphics and Application. 1999; 19(5): 0020-21.
14. Wongsuphasawat K, Guerra Gómez JA, Plaisant C, Wang TD, Taieb-Maimon M, Shneiderman B. LifeFlow: visualizing an overview of event sequences. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2011; 1747-1756.
15. Monroe M, Lan R, Lee H, Plaisant C, Shneiderman B. Temporal event sequence simplification. IEEE Transactions on Visualization and Computer Graphics. 2013; 19(12): 2227-2236.
16. Lin J, Keogh E, Lonardi S, Lankford JP, Nystrom DM. VizTree: a tool for visually mining and monitoring massive time series databases. In Proceedings of the Thirtieth international conference on Very large data bases. 2004; 30: 1269-1272.
17. Havre S, Hetzler E, Whitney P, Nowell L. (2002). Themeriver: Visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics. 2002; 8(1): 9-20.
18. Zheng K, Haftel HM, Hirschl RB, O'Reilly M, Hanauer DA. Quantifying the impact of health IT implementations on clinical workflow: a new methodological perspective. Journal of the American Medical Informatics Association. 2010; 17(4): 454-461.
19. Li C, Biswas G. A Bayesian Approach to Temporal Data Clustering using Hidden Markov Models. In ICML. 2000; 543-550.
20. Lin J, Keogh E, Lonardi S, Chiu B. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. 2003.
21. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009; 10(3), R25.
22. Keogh E, Lonardi S, Chiu BYC. Finding surprising patterns in a time series database in linear time and space. In Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining. 2002; 550-556.
23. Tseng VS, Lee CH. Effective temporal data classification by integrating sequential pattern mining and probabilistic induction. Expert Systems with Applications. 2009; 36(5): 9524-9532.
24. Cheng HC, Plaisant C, Shneiderman B. Identifying and Measuring Associations of Temporal Events. HCIL-2012-05.
25. Fong A, Meadors M, Batta N, Nitzber M, Hettinger AZ, Ratwani R. Identifying Interruption Clusters in the Emergency Department, In Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2014; (in press).

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.

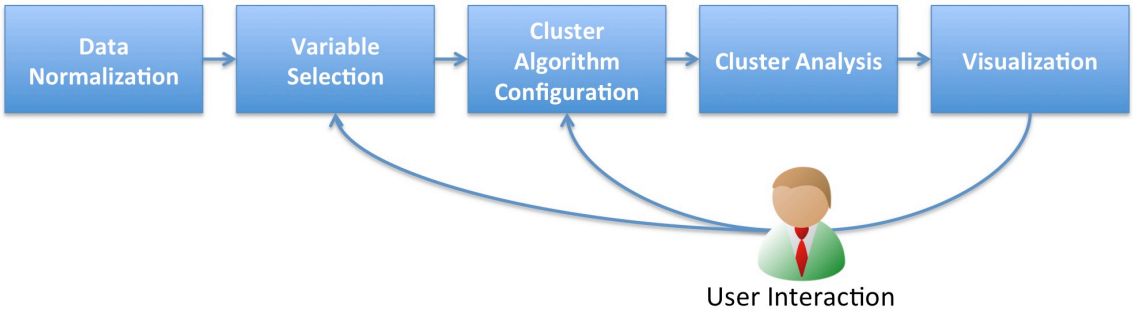


Figure 1. Our method begins with raw survey data which is then normalized and prepared for cluster analysis. Users can optionally select a subset of survey variables (all variables are used by default) and cluster algorithm parameters to use for computing the clusters at each time step. The clusters 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.

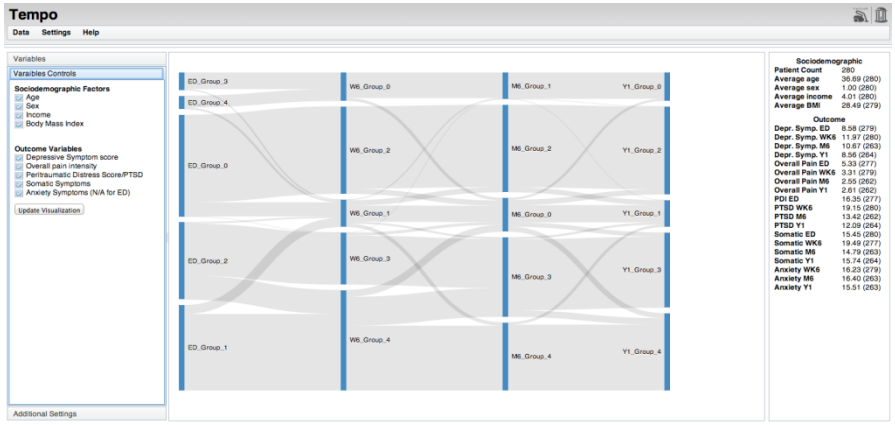


Figure 2. A screen capture of our prototype implementation applied the study data. The four vertical lines of blue rectangles correspond to the four time steps in the data: ED, W6, M6, and Y1. The left sidebar shows the systems variable selection controls while the right sidebar shows detailed statistics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

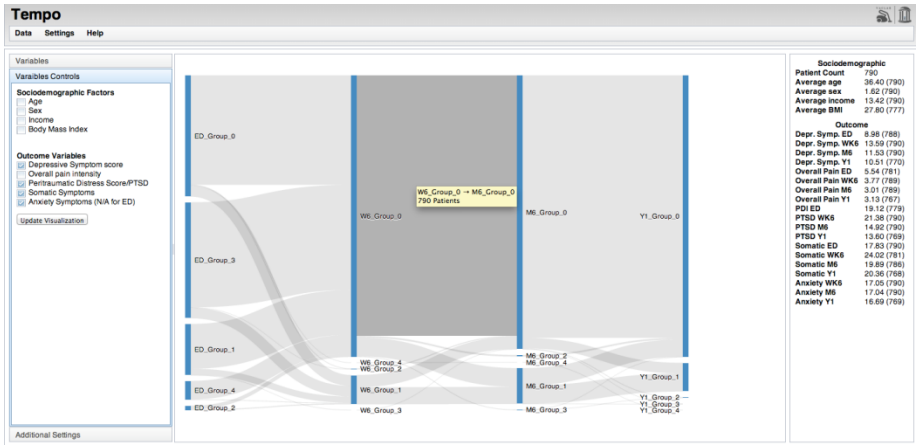


Figure 3. An alternative view of the same data being visualized in Figure 2. The differences in pathways in the visualization are caused directly by the user's interaction with the variable controls panel.

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).

Visualizations of Inter-Observer Reliability Assessments in Time Motion Studies: Facilitating Observers' Training.

Marcelo Lopetegui[1,2], MD, MS, Po-Yin Yen[1], RN, PhD, Alejandro Mauro[2], MD, Barbara Lara[1], MD, MPH, Philip Payne[1], PhD

[1] Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA.
[2] Clínica Alemana de Santiago, Facultad de Medicina Clínica Alemana, Universidad del Desarrollo, Santiago, Chile.

Abstract

Time Motion Studies are used to collect the quantitative data required to study clinical workflow and provide evidence for decision makers in healthcare. These are time and resource intensive studies, partially due to the extensive training process required to achieve acceptable reliability among observers. We postulated that the training process could be optimized by providing visualizations accompanying the commonly cited statistics, informing researchers on directed actions to improve the observers' reliability based on these visualizations. A prototype was developed and used in a workflow time study. Researchers perceived the visualizations as a substantial contribution to the observers' training process.

INTRODUCTION

Clinical workflow is a current topic of research addressing some of the most critical issues in healthcare delivery, including patient safety and quality of care by detecting medication errors[1,2] and assessing timeliness of treatments[3,4] and procedures[5,6], also focusing on productivity[7,8], efficiency[9,10] and optimization of clinicians' workload[11,12]. In order to collect the necessary data and information to assess these issues, hospital administrators and clinical workflow researchers rely on mixed methods. To *understand*, they use qualitative approaches such as ethnographic studies and interviews[13]. To *measure*, they rely on quantitative time and motion measurement methods borrowed from the industrial engineering arena[14,15], generally known as Time Motion Studies (TMS).

Among the techniques available to conduct TMS, "workflow time study" is the most used[16]. In this variation of TMS, observers continuously follow a subject for a predefined period of time and record tasks as they occur, producing a data schema of time-stamped sequences of tasks[12,17]. This technique allows observers to track unexpected instances of tasks, accounting for task fragmentation, interruptions, and the real-world variability of clinical workflow. However, unlike other TMS techniques, workflow time studies substantially increase the burden on observers, raising concern over observers' reliability. Although every data collection method requiring a human interface should include a reliability assessment, this is not systematically reported in TMS due to a lack of a standards to conduct Inter-Observer Reliability Assessments in these scenarios (I.O.R.A.) [18].

IORA in workflow time studies is usually conducted by having two independent observers follow a common subject at the same time, capturing data separately, and then comparing the recorded data [**Figure 1**]. The myriad of statistics used for these analyses include the Kappa coefficient, the Bland-Atman plot and Limit-of-Agreement estimates, Pearson's correlation and Intra-class correlation[18]. Although these methods only assess one dimension of the agreement (either duration or

Figure 1: Example of how IORA is conducted in workflow time studies. Two observers follow the same subject and capture data independently, to later assess their agreement in naming and/or timing the recognized tasks. *Modified from spotmatikphoto © 123RF.com, standard license*.

naming of tasks), they provide an interpretable scalar value that authors use to report reliability (with arbitrarily cut points for acceptable reliability).

Regardless of the method reported, researchers attempt to achieve and maintain a minimum acceptable IORA in order to assert valid data for their analyses. Thus, these IORA play a crucial role in the observer's training, by reassuring the investigator that his observers are consistent with each other and ready to capture valid data.

Conducting IORA and training observers is a very resource intensive activity, devoting several hours to achieve an acceptable reliability[19], guided mainly by the IORA scores produced by the statistics above mentioned. We hypothesize that the training of the observers' could be improved and facilitated by providing investigators with visualizations describing how those results were obtained. On top of reading a score that informs "how" observer are performing, we envision visualizations providing information on "why" observers achieve that score, thus empowering the investigator with knowledge on directed actions to improve the reliability.

In this project, we aimed to design visualizations to complement the most used statistics in IORA, implement them onto a time capture tool broadly used to conduct TMS[20] and evaluate the perceived impact of these visualizations on the observers' training phase of a TMS.

METHODS

Visualizations.

We focused on two of the most cited statistics used to report IORA[18]: The Kappa coefficient and Pearson's correlation to report agreement in naming and in timing respectively. The Kappa coefficient measures pairwise agreement among a set of coders making categorical judgments, correcting for expected change agreement[21], and is usually reported as a float between 0 and 1 (predefined cut points to interpret the agreement are generally accepted[22]). Pearson's correlation is also informed as a float, between -1 and 1 (interpreted as negative and positive correlation respectively, or 0 for no correlation). We focused our design efforts on providing the investigators with visualizations on the contingency tables used to calculate kappa and the plot to visualize deviations from the best-fit line for Pearson's correlation, informing them on agreement on each pair of tasks. We used a user centered design approach working closely with two experienced researchers in TMS. The visualization were iteratively improved and validated throughout the development cycle.

The Software.

The Time Capture Tool (TimeCaT) is a comprehensive, flexible, and user-centered web application developed by the department of biomedical informatics at The Ohio State University to support data capture for TMS[20]. It provides a friendly touch-time-stamp interface: the user simply clicks a button with the loaded task-name, and a time-stamp is created. TimeCaT is accessible at http://www.timecat.org and available at no cost to non-profit researchers. Although the tool is not open source, we participate in the development and have knowledge of the system, architecture, and database schemas. We implemented the visualized IORA on a Beta release of TimeCaT.

The Workflow Time Study

The visualizations were implemented in TimeCaT and used for a workflow time study conducted in an outpatient clinic of The Ohio State University Wexner Medical Center. The main research question focused on changes in workflow and impact on duration of specific sections of patient encounters. The workflow time study principal investigator had previous experience conducting these studies, and was a familiar user of TimeCaT. A medical doctor with previous experience in conducting continuous observation and familiar with the outpatient clinic workflow served as the gold standard observer, training six nursing students. We invited them to use the BETA release of TimeCaT which included our visualizations for the IORA module. They agreed to participate in the evaluation of the IORA visualizations, and included us in their IRB protocol due to the nature of the data being collected. A total of twenty four IORA sessions were conducted (4 observations per each trainee against the gold standard observer).

RESULTS

Visualizations

We designed and implemented three visualizations in the IORA module on TimeCaT: a side by side visualization of both observations, contingency tables for kappa calculations, and a scatter plot for Pearson's correlation coefficient. In order to maximize the ease of interpretation and complementariness of the visualizations, different colors were arbitrarily assigned to task names and were maintained consistent within the report, helping to identify tasks across visualizations. The reports were made available immediately after the observations were completed, providing on-site feedback before performing further IORA sessions. Given the web-based nature of the application, the principal investigator could simultaneously assess the performance remotely and provide guidance if required.

Side by side visualization of an IORA session. Based on the asynchronous sample report of Mache's tool[23], we designed side by side visualizations of each IORA session, representing the gold standard next to the trainee, and added a zoom functionality. In this visualization, each observation is displayed vertically from start to end (top to bottom), and each sequential task is represented of a different color (representing different tasks names), while the duration of each task is represented by the height of the block **[Figure 2]**.

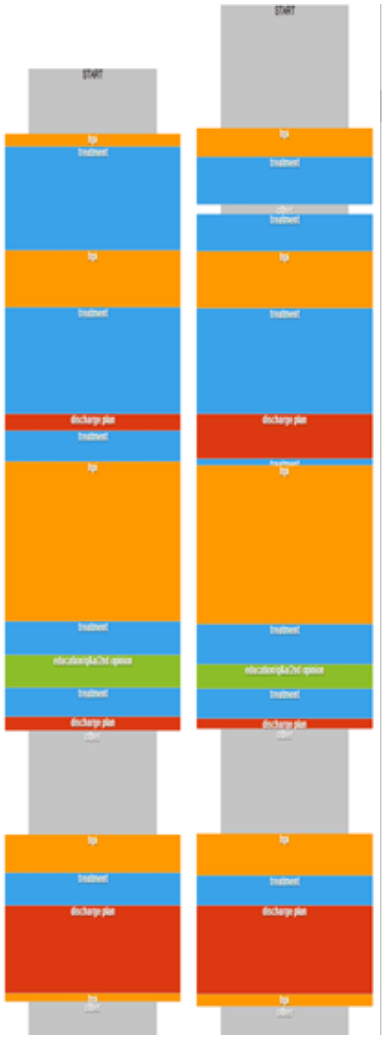


Figure 2: side by side visualization of an IORA session. Each column represents one observation: the gold standard observer on the left, and the trainee on the right. The investigator can easily identify any major disagreements at a glance.

Contingency tables for kappa calculations. Kappa's contingency tables were created, and iteratively refined from a raw model to an easy to interpret graphic. Incremental improvements resulting from the validation included getting rid of empty cells content (removing the zeros), fixing cell width by verticalizing the table headers, coloring the agreements in green and the disagreements in red, and providing a hover effect to easily identify row/columns intersections **[Figure 3]**.

Scatter plot for Pearson's correlation coefficient. Since the desired outcome of the visualizations was not to re-display the Pearson result (assessing the slope of the best fit line), in lieu of plotting the best fit line for the paired data regarding tasks duration, we created a visualization showing the slope of one (perfect positive agreement). Thus, the investigators could easily assess which points fell below the line, representing pairs where the trainee recorded a shorter duration than expected, and which points fell above the line, representing pairs where the trainee recorded a longer duration than expected **[Figure 4]**.

Feedback from the users

The composite IORA resulted in a solid foundation to guide observers' training. Both the principal investigator and the gold standard observer found the visualizations to be easy to interpret, identifying the side-by-side visualizations as the most useful component to train the observers. They also highlighted the ability to assess the

	HPI	treatment	other	Discharge	Education
Hpi	290	5	0	1	0
Treatment	18	301	10	30	1
Other	12	0	896	0	0
Discharge	0	2	2	123	0
Education	0	10	0	0	25

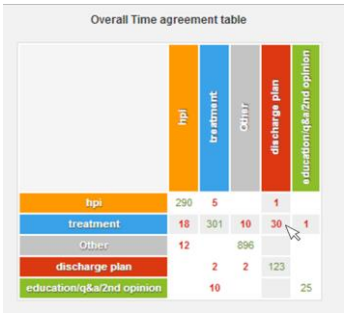


Figure 3: Contingency tables for Kappa calculations. On top, the original table resulting from the IORA session. Bottom, the optimized visualization of the same table, removing empty cells, contrasting agreements v/s disagreements, maintaining color scheme for tasks, and providing a hover effect for rows and columns for ease of interpretation.

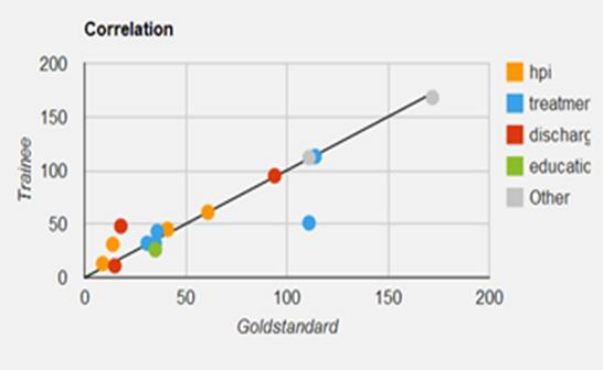


Figure 4: Scatter plot representing the concordance between the duration of each task captured by the gold standard and the trainee. Everything below the black diagonal represents a task captured by the trainee that is shorter than the gold standard. Anything above represents a task captured by the trainee that is longer than the gold standard.

reliability on-site in real time as a major advantage compared to any other time capture tool, having complete reports to discuss ways to improve the performance in the field, producing a clearer feedback for the trainees. Based on the visualizations they were able to discuss errors such as the granularity of the tasks being recorded (the trainee capturing more tasks than the gold standard), errors in naming tasks (the trainee naming a task differently) and errors in timing tasks (the trainee being late or early at recording the data), all of which were either unavailable, or required a long, asynchronous and manual analysis of the data.

DISCUSSION

Our visualization module was perceived as an optimization of the current training process by researchers conducting workflow time studies. The clear information allowed them to take directed corrective actions, previously unavailable. Although the side-by-side visualization was first introduced by Mache's tool[23], the web-application nature of TimeCaT provided a unique advantage: the gold-standard can assess the IORA session on-site, immediately after the observation is completed. Thus, he could provide immediate feedback and take corrective actions in the field, where it seemed to be more helpful for the trainees. Finally, the new visualizations for Kappa and Pearson proved to be a helpful feature for more in-deep assessments of click-accuracy, both in naming tasks and timing tasks. An optimized training phase could reduce costs of conducting TMS, and moreover, lead to increased reliability among the observers, which in turn produces more valid data to analyze and reason upon.

The concerns on the suitability of each of the statistical tests reported in IORA for TMS, as well as apprehensions on the data transformation required to implement them, are beyond the scope of this article. We acknowledged the lack of standards on how to conduct IORA, and thus based our visualization on two of the most used statistics in order to provide researchers with more information to optimize the training of their observers, given the current practices. We are not promoting nor validating the use of these statistics. Our team is concurrently working on a new composite IORA protocol to empower workflow researchers with a standardized and comprehensive method for validating observers' reliability and, in turn, the validity and representativeness of the data collected (unpublished work).

Limitations

Our evaluation only consisted on a single workflow time study, and assessed the perception of two users. Given our positive findings, we included the module as part of the latest stable TimeCaT release and expect to collect broader feedback and run formal usability evaluations in the future.

CONCLUSION

The IORA visualizations proved to be perceived as a substantial contribution to observers' training. By optimizing the training process, they could reduce the time and resources required to train observers, and even contribute to achieve better IORA scores, which gets translated on to more valid data from these observations. The proposed visualizations for IORA are freely available for researchers using TimeCaT, accessible at www.timecat.org.

REFERENCES

1. Barker KN, Flynn E a, Pepper G a. Observation method of detecting medication errors. *Am J Health Syst Pharm*. 2002;59(23):2314-6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12489369.

2. Keohane CA, Bane AD, Featherstone E, et al. Quantifying nursing workflow in medication administration. *J Nurs Adm*. 2008;38(1):19-26. doi:10.1097/01.NNA.0000295628.87968.bc.

3. Rebmann T, Clements BW, Bailey JA, Evans RG. Organophosphate antidote auto-injectors vs. traditional administration: a time motion study. *J Emerg Med*. 2009;37(2):139-43. doi:10.1016/j.jemermed.2007.09.043.

4. Wolfrum S, Pierau C, Radke PW, Schunkert H, Kurowski V. Mild therapeutic hypothermia in patients after out-of-hospital cardiac arrest due to acute ST-segment elevation myocardial infarction undergoing immediate percutaneous coronary intervention. *Crit Care Med*. 2008;36(6):1780-6. doi:10.1097/CCM.0b013e31817437ca.

5. Wiese CHR, Bartels U, Bergmann A, Bergmann I, Bahr J, Graf BM. Using a laryngeal tube during cardiac arrest reduces "no flow time" in a manikin study: a comparison between laryngeal tube and endotracheal tube. *Wien Klin Wochenschr*. 2008;120(7-8):217-23. doi:10.1007/s00508-008-0953-1.

6. Zalaudek I, Kittler H, Marghoob AA, et al. Time required for a complete skin examination with and without dermoscopy: a prospective, randomized multicenter study. *Arch Dermatol*. 2008;144(4):509-13. doi:10.1001/archderm.144.4.509.

7. Were MC, Sutherland JM, Bwana M, Ssali J, Emenyonu N, Tierney WM. Patterns of care in two HIV continuity clinics in Uganda, Africa: a time-motion study. *AIDS Care*. 2008;20(6):677-82. doi:10.1080/09540120701687067.

8. Schiller B, Doss S, DE Cock E, Del Aguila MA, Nissenson AR. Costs of managing anemia with erythropoiesis-stimulating agents during hemodialysis: a time and motion study. *Hemodial Int*. 2008;12(4):441-9. doi:10.1111/j.1542-4758.2008.00308.x.

9. Fisher J, Lotery H, Henderson C. Time in motion--testing efficiency in the dermatology procedure setting. *Dermatol Surg*. 2009;35(3):437-44; discussion 445. doi:10.1111/j.1524-4725.2009.01076.x.

10. Amusan AA, Tongen S, Speedie SM, Mellin A. A time-motion study to evaluate the impact of EMR and CPOE implementation on physician efficiency. *J Healthc Inf Manag*. 2008;22(4):31-7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19267017. Accessed September 27, 2011.

11. Trotter MJ, Larsen ET, Tait N, Wright JR. Time study of clinical and nonclinical workload in pathology and laboratory medicine. *Am J Clin Pathol*. 2009;131(6):759-67. doi:10.1309/AJCP8SKO6BUJQXHD.

12. Westbrook JI, Ampt A, Kearney L, Rob MI. All in a day's work: an observational study to quantify how and with whom doctors on hospital wards spend their time. *Med J Aust*. 2008;188(9):506-9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18459920. Accessed July 19, 2011.

13. Malhotra S, Jordan D, Shortliffe E, Patel VL. Workflow modeling in critical care: piecing together your own puzzle. *J Biomed Inform*. 2007;40(2):81-92. doi:10.1016/j.jbi.2006.06.002.

14. Clinical rounds: Nursing activities: how paperwork eats up your time. *Plast Surg Nurs*. 30(2):123. doi:10.1097/PSN.0b013e3181ebc7b7.

15. Hendrich A, Chow MP, Skierczynski BA, Lu Z. A 36-hospital time and motion study: how do medical-surgical nurses spend their time? *Perm J*. 2008;12(3):25-34. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037121&tool=pmcentrez&rendertype=abstract. Accessed November 24, 2011.

16. Lopetegui M, Yen P-Y, Lai A, Jeffries J, Embi P, Payne P. Time Motion Studies in Healthcare: What are we talking about? *J Biomed Inform*. 2014. doi:10.1016/j.jbi.2014.02.017.

17. Mache S, Busch D, Vitzthum K, Kusma B, Klapp BF, Groneberg DA. Cardiologists' workflow in small to medium-sized German hospitals: an observational work analysis. *J Cardiovasc Med (Hagerstown)*. 2011;12(7):475-81. doi:10.2459/JCM.0b013e328347db8f.

18. Lopetegui MA, Bai S, Yen P-Y, Lai A, Embi P, Payne PRO. Inter-observer reliability assessments in time motion studies: the foundation for meaningful clinical workflow analysis. *AMIA Annu Symp Proc*. 2013;2013:889-96. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3900222&tool=pmcentrez&rendertype=abstract. Accessed May 27, 2014.

19. Lara B, Meara A, Ardoin S, Embi P, Yen P. Challenges Faced When Designing and Conducting Time Motion Studies in Health Care Environments. *AMIA Annu Symp Proc*. 2014.

20. Lopetegui M, Yen P, Lai AM, Embi PJ, Payne PRO. Time Capture Tool (TimeCaT): Development of a Comprehensive Application to Support Data Capture for Time Motion Studies . *AMIA Annu Symp Proc*. 2012.

21. Carletta J. Assessing agreement on classification tasks: the kappa statistic. *Comput Linguist*. 1996;22(2):249-254. Available at: http://dl.acm.org/citation.cfm?id=230386.230390. Accessed February 18, 2014.

22. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med*. 2005;37(5):360-3. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15883903. Accessed February 18, 2014.

23. Mache S, Scutaru C, Vitzthum K, et al. Development and evaluation of a computer-based medical work assessment programme. *J Occup Med Toxicol*. 2008;3:35. doi:10.1186/1745-6673-3-35.

Multivariate Visualization of System-Wide National Health Service Data Using Radial Coordinates

David Borland[1], Vivian L. West[2], W. Ed Hammond[2]
[1]RENCI, The University of North Carolina at Chapel Hill, Chapel Hill, NC;
[2]Duke Center for Health Informatics, Duke University, Durham, NC

Abstract

We present radial coordinates, a multivariate visualization technique based on parallel coordinates. The visualization contains a number of features driven by the needs of health-related data analysis, such as integrating categorical and numeric data, and comparing user-selected subpopulations via ribbon rendering. We illustrate the utility of radial coordinates by exploring primary care trust (PCT) and practice-level data from the United Kingdom's National Health Service, using three examples: lung cancer rates among PCTs, various cancer rates among only London suburb PCTs, and medical problem prevalence among over 1500 London practices.

Introduction

With the ever-increasing size and number of health-related datasets, new analytical tools are becoming necessary to enable enhanced understanding of the vast amount of information contained within. Visualization leverages the power of the human visual system to reveal patterns and relationships in data by mapping the data to visually salient features.

One of the challenges for visualization of health-related data is the desire to incorporate data of many types (e.g. lab results, demographics, medications, vital signs, and genomic data) from various sources. We have developed a multivariate visualization technique, radial coordinates, that enables visual analysis of a wide range of health-related datasets and handles both numeric and categorical data (Figure 1). Radial coordinates facilitates the interactive exploration of datasets to reveal patterns in the data, discover relationships between variables, and compare user-defined subpopulations. In this manner we support the pursuit of hypothesis formations that can elicit further inquiry and lead to new knowledge.

An overview of an initial radial coordinates prototype applied to query data was given previously.[1] In this paper we provide a more in-depth description of the various features of a new implementation, which includes several new features, and discuss its application to primary care trust (PCT) and practice-level data from the National Health Service (NHS) in the United Kingdom (UK). We present three examples illustrating the use of radial coordinates to explore the NHS data: lung cancer rates among PCTs, a comparison of various cancer rates among London suburb PCTs, and medical problem prevalence among over 1500 London practices.

Previous Work

Our visualization is based largely on parallel coordinates, a multivariate visualization technique which represents each dimension as a parallel axis, and each data entity as a line connecting the entity's value at each axis.[2,3] Non-parallel arrangements of axes have also been investigated.[4] Our radial coordinates arrangement differs in that the radial layout maintains a square aspect ratio even with many axes, and enables utilization of the space in the center of the radial layout. Parallel coordinates have been combined with various other visualization techniques[5-7], including direct integration of scatter plots.[8,9] In our visualization we include a scatter plot based on the first two principal components to enhance the ability to find clusters in high-dimensional data in an intuitive manner (Figure 1a). Future work will explore combinations with other techniques. We also incorporate chords representing the correlations between axes in a manner similar to Circos.[10] Extensions to parallel coordinates for incorporating categorical data include parallel sets[11] and hammock plots.[12] Both represent multiple data points as paths between axes, with the number of data points encoded as path width. Our curve spreading technique incorporates categorical and continuous data while still enabling the visualization of individual data points (Figure 2). Various techniques have been developed to combine multiple data points to enhance the understanding of large datasets[13,14] and observe clusters via edge bundling techniques.[15,16] Our ribbon rendering technique enables enhanced visualization of user-selected data points, including overlaying information of statistical data (median value and quartile ranges) of interest to the health-care community (Figure 1b). Axis ordering is an important element of parallel coordinates visualizations, as it is typically easier to notice relationships between variables with adjacent axes.[17-19] We employ a correlation-based clustering technique and also introduce dynamic reordering of categorical axis values to cluster similar values based on user-defined selections (Figures 3c, 3d).

Methods

Data

PCTs, abolished in 2013 due to NHS reorganization, were regional administrative bodies in the UK responsible for commissioning health services from providers and providing community health services. Here we investigate 26 variables measuring various health and socioeconomic factors for 147 of the 152 PCTs in England (five were removed due to missing data). Health factors include cancer rates, drug prescription rates, and factors related to diabetes prevalence and treatment. Socioeconomic factors include socioeconomic deprivation, economic output, geographic region, and local region classification (e.g. *Manufacturing Towns* and *Coastal and Countryside*) from the Office for National Statistics (ONS).

We also demonstrate our visualization with data showing the prevalence of a number of medical problems (e.g. diabetes, dementia, and obesity) in the 1504 practices in the London strategic health authority (SHA). There were ten SHAs in England from 2006-2013.

Visualization

The radial coordinates visualization, implemented using the D3 JavaScript library[20], represents each variable in a multivariate dataset by an axis, with the axes arranged radially around a circle. Each individual entity is represented by a curve that connects the value of that entity at each axis. Figure 1 gives an example applied to PCT data, with four PCT curves highlighted in red by the user.

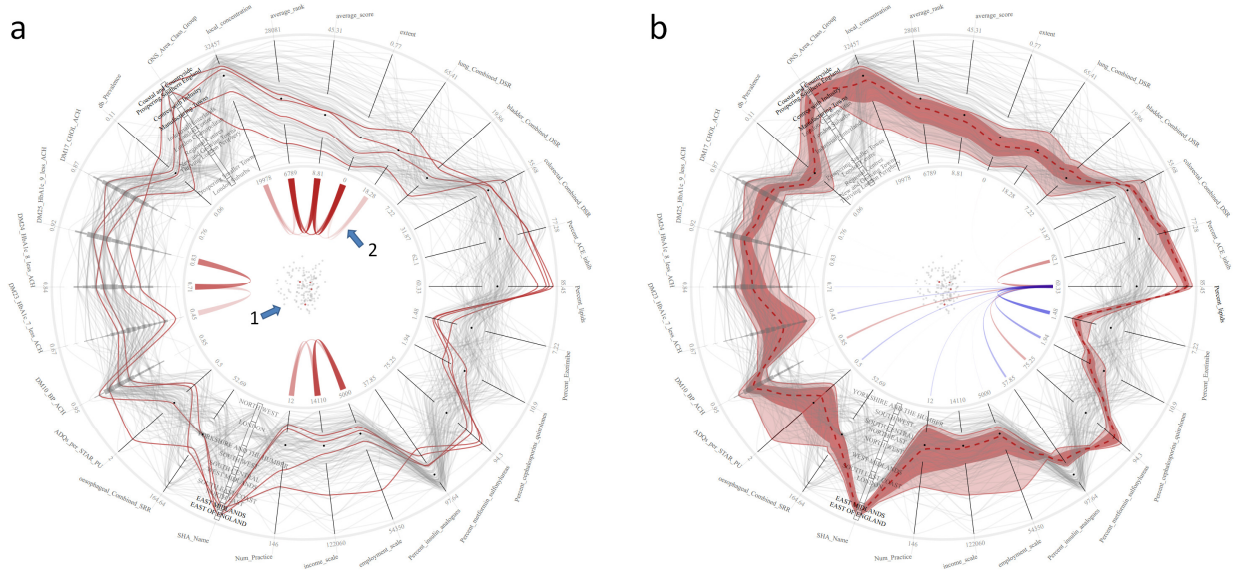


Figure 1. Radial coordinates visualizations of NHS PCT data. User-highlighted curves (red) enable the comparison of four PCTs across multiple variables (a). A linked scatterplot of the first two principal components can help show clusters in high-dimensions (a1). Chords connecting axes represent correlations (positive: red, negative: blue) above a user-defined threshold (a2). Ribbon rendering enables a simplified representation of user-defined subpopulations, displaying the data range optionally overlaid with median value and inner quartile ranges (b). Mouse over of an axis shows all correlations with that axis, regardless of user-defined threshold (b).

User selection of individual curves enables a visual comparison of how different entities relate across the various axes. A radial layout elegantly handles large numbers of axes while maintaining a square aspect ratio, also enabling the use of the center of the layout for supplemental visualizations, such as axis correlation chords and a scatterplot of the first two principal components (Figure 1a). Ribbon rendering uses a sliding window algorithm to draw the area between the innermost and outermost boundary of selected curves in a semi-transparent solid color, making it easier to see the spread of each subpopulation. An optional summary statistic overlay shows the inner quartile range and median value of each subpopulation (Figure 1b). Other visualization features include data-type dependent axis distribution visualizations and curve spreading for categorical and discrete data (Figure 2).

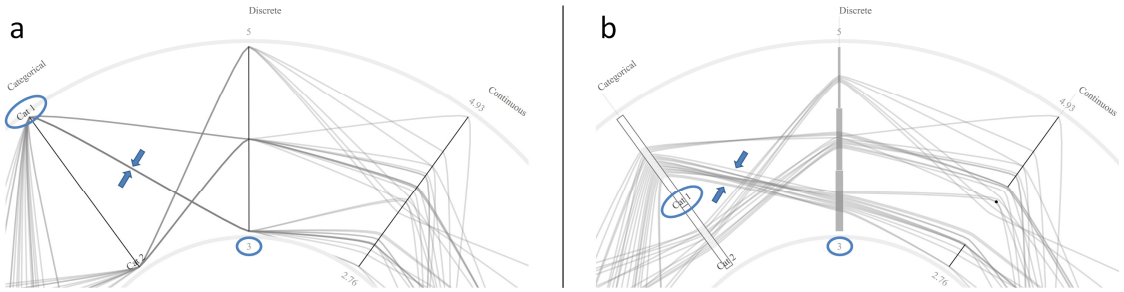


Figure 2. A sample data set without (a) and with (b) data-type dependent axis distribution visualizations and curve spreading. Axis distribution visualizations represent categorical axes as a stacked bar chart, discrete numeric axes as a histogram, and continuous numeric axes as a quartile plot[21], enabling rapid evaluation of the data type and overall distribution of the data for each axis. Curve spreading for categorical and discrete axes enables improved visualization of individual curves and clusters of curves, such as the number of data points with a Categorical value of Cat 1 and a Discrete value of three (highlighted in blue).

Results

Lung Cancer Prevalence

In Figure 3 the user has clicked on the lung cancer rate axis (*lung_Combined_DSR*), causing PCTs in the upper quartile of lung cancer rate to be automatically colored red, and the lower quartile blue. High and low lung cancer rates can now be compared across all dimensions in the data (Figure 3a). In the upper portion of the visualization it is apparent that PCTs with high and low lung cancer rates also tend to have high and low values for *extent*, *average_score*, *average_rank*, and *local_concentration* (also indicated by the correlation chords connecting these axes), which represent measures of social deprivation (poverty rate, socioeconomic status, etc.)

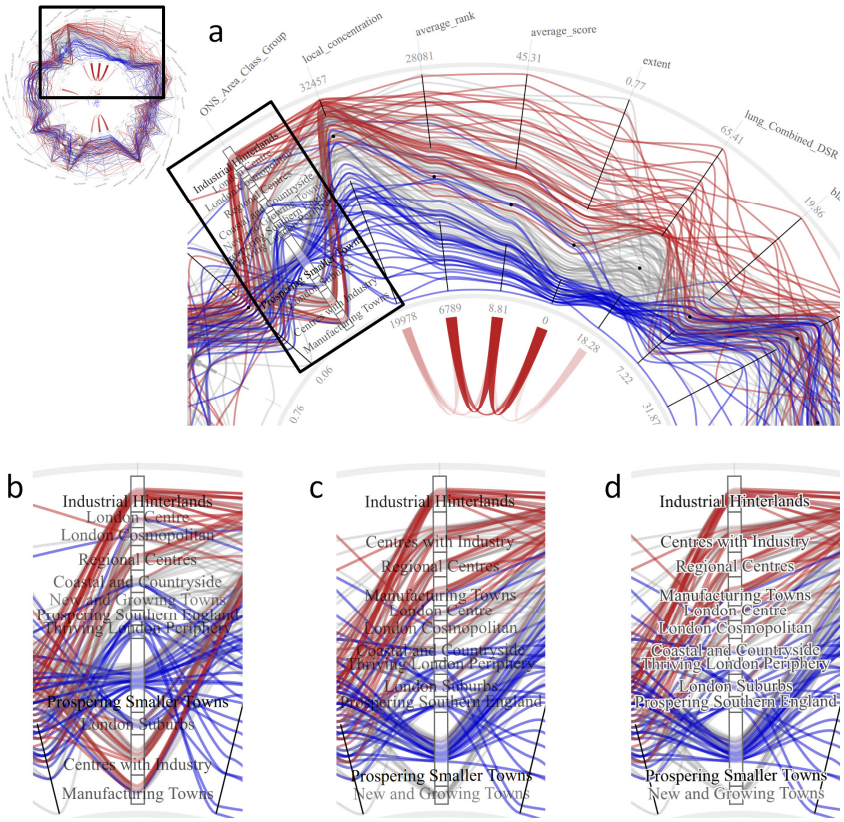


Figure 3. Visualization of lung cancer rates (red = upper quartile, blue = lower quartile) in 147 primary care trusts (PCTs) in the UK. High and low lung cancer rates tend to cluster based on regional classification (b), made clearer with automatic categorical axis reordering to cluster similar regions (c, d).

The red and blue curves also form clusters on the *ONS_Area_Class_Group* axis, a local region categorization from the ONS. Investigating this axis (Figures 3b-d) shows that *Industrial Hinterlands*, *Centres with Industry*, *Regional Centers*, and *Manufacturing Towns* all have high lung cancer rates, whereas *Prospering Smaller Towns*, *Prospering Southern England*, *London Suburbs*, and *Thriving London Periphery* all have low lung cancer rates. The discovery of such relationships via exploring the data visually drives the formation of causal hypotheses (e.g. pollution levels or smoking prevalence), which can be investigated further.

London Suburb Comparison

In Figure 4a a single PCT, Harrow, was seen to have the lowest lung, bladder, and colorectal cancer rates compared to all other PCTs, and has been highlighted in red. Harrow is classified as a *London Suburb*, so in Figure 4b the user has highlighted the other London suburbs in blue for comparison, made easier in Figure 4c via ribbon rendering. Harrow is shown to have a much higher value for the *oesophageal_Combined_SRR* axis, and thus a much higher esophageal cancer rate, than the other London suburbs, which are almost all in the lower quartile. This visualization raises the question of why Harrow has such a disparity in the rates of different cancers.

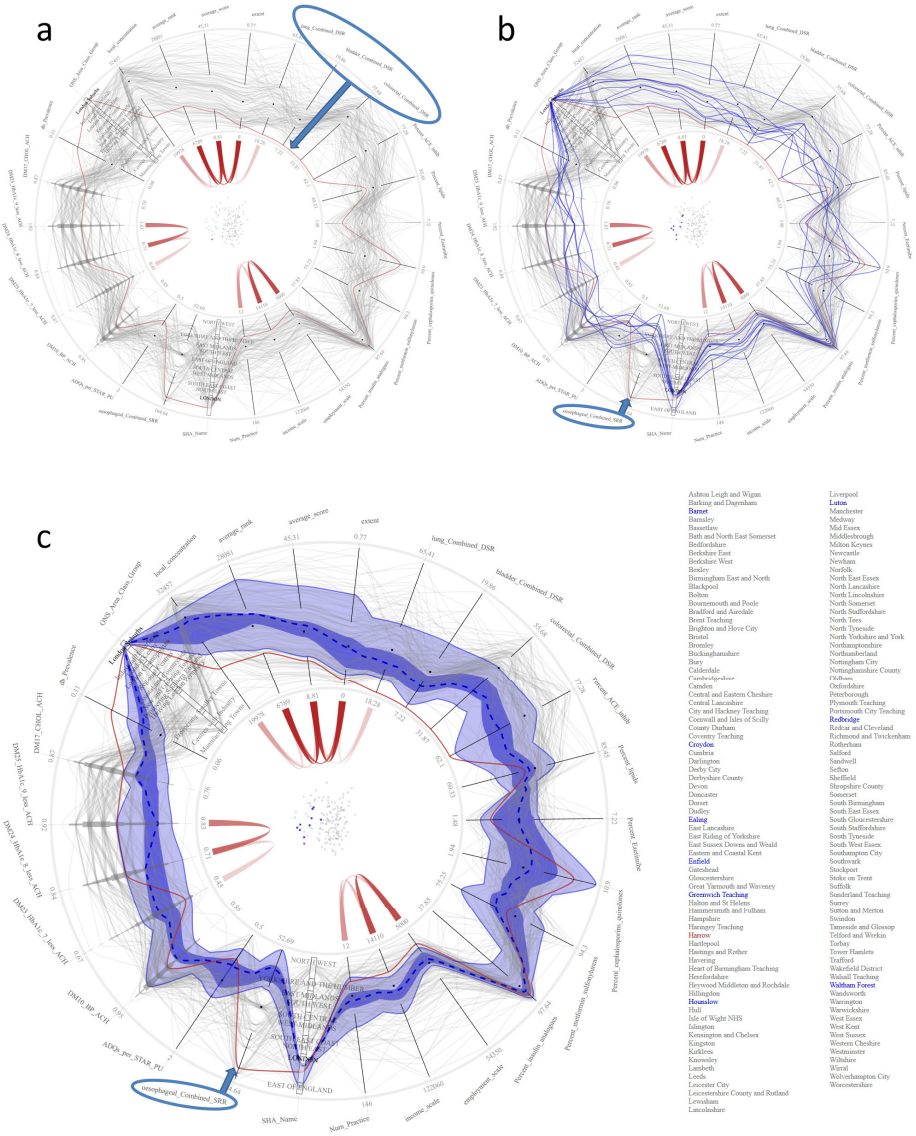


Figure 4. The Harrow PCT (red) has the lowest lung, bladder, and colorectal cancer rates (circled) among all 147 PCTs in the NHS dataset (a). Comparing Harrow to the other London suburbs (blue) reveals that Harrow has a much higher esophageal cancer rate (circled) than the other suburbs (b). Ribbon rendering makes it easier to visually compare Harrow with the other London suburbs (c).

According to the 2011 Census[22] Harrow is very diverse, with 63.8% of its population from the Black and Minority Ethnic communities, including the highest concentration of Sri Lankan Tamils and Gujarati Hindus in the UK and Ireland. India is known to have relatively low cancer rates in general, but some of the highest rates for oral and esophageal cancers in the world[23], which may help explain this phenomenon. Although further analysis is necessary, this example shows the utility of radial coordinates and ribbon rendering to compare subpopulations.

Practice-Level Data

Figures 8a and 8b show the prevalence of various medical problems (e.g. diabetes, dementia, and obesity) in the 1504 practices in the London strategic health authority (SHA). Figure 8a highlights in red two practices that appear to be outliers in the PCA scatterplot. Ribbon rendering makes apparent that they have the two highest prevalences for 12 of the 21 medical problems represented in the data. Figure 8b applies ribbon rendering to the remaining 1502 practices, making it easier to compare maximum and minimum values of medial problem rates for the two subpopulations.

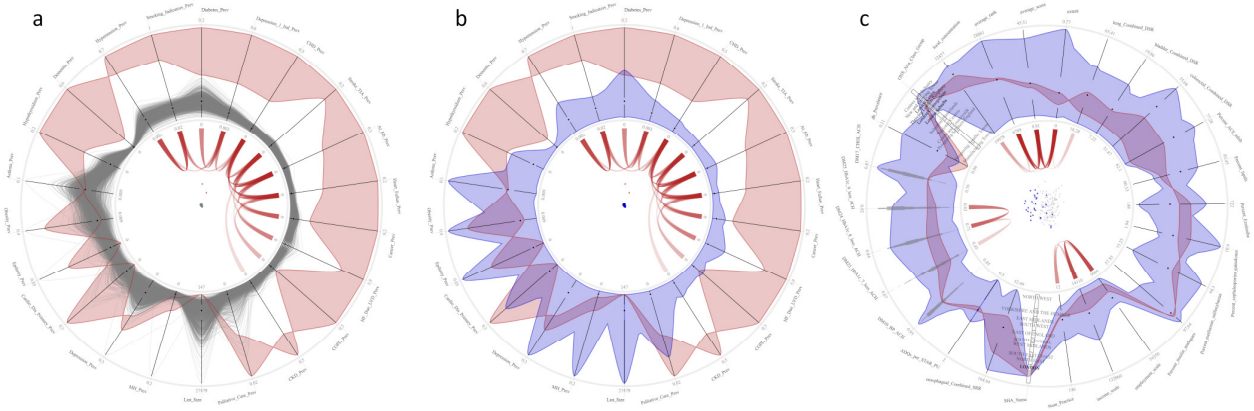


Figure 8. Two out of the 1504 practices in the London SHA, highlighted in red, have the two highest prevalences for 12 of the 21 medical problems represented in the NHS practice-level data (a and b). Comparing the PCTs containing these practices (red) to all other London PCTs (blue) does not reveal any major differences (c).

The two practices highlighted in red are Royal Hospital Chelsea in the Kensington and Chelsea PCT, and Nightingale House in the Wandsworth PCT. Because these two practices stood out so dramatically in the practice-level data, the user performed a PCT-level comparison of all London PCTs (Figure 8c). Interestingly, the Kensington and Chelsea and the Nightingale House PCTs (red) do not appear very different when compared to the other London PCTs (blue). Further research determined that Royal Hospital Chelsea is a retirement and nursing home for British soldiers and Nightingale House is a nursing home for the Jewish community that specializes in dementia, which may explain the high prevalence of problems such as dementia, hypertension, stroke, heart failure, and cancer in these two practices.

Conclusion

We have presented radial coordinates, a multivariate visualization technique based on parallel coordinates that incorporates features, such as per-axis population distribution visualizations based on data type (continuous, discrete, and categorical), direct visualization of correlations between variables, curve spreading for discrete and categorical data, visualization of summary statistics for user-selected subpopulations via ribbon rendering, and automatic reordering of categorical values based on user selection, driven by the needs of health-related data visualization.

We have applied radial coordinates to data from the UK's NHS at both the PCT and individual practice levels. Visualization of lung cancer rates among PCTs discovered possible relationships among lung cancer rate, socioeconomic factors, and regional classification. A comparison of London suburb PCTs revealed a potentially interesting PCT with a much higher esophageal cancer rate than other similar PCTs. Visualizing medical problem prevalence among over 1500 London practices showed two practices that have much higher rates of many medical problems. These examples illustrate the utility of the combination of visualization techniques embodied in our radial coordinates tool, and underline the need for further research in the use of visualization to aid in the analysis of complicated health-related datasets.

Acknowledgments

NHS and other UK data were made available courtesy of the BT Health Cloud. This work is supported 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. West V, Borland D, Hammond WE. Visualization of EHR and Health Related Data for Information Discovery. In Proceedings of the 2013 AMIA Workshop on Visual Analytics in Healthcare. November 2013.
2. Gannet H. General summary, showing the rank of states, by ratios. 1880.
3. Inselberg A. The plane with parallel coordinates. Visual Computer. 1985;1(4):69-91.
4. Tominiski C, Schumann H. An event-based approach to visualization. In Proceedings of the Eighth International Conference on Information Visualization (IV'04). July 2004;101-107.
5. Rodrigues Jr. JF, Traina AJM, Traina Jr. C. Frequency plot and relevance plot to enhance visual data exploration. In Proceedings of the XVI Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI'03). 2003;117-124.
6. Edsall RM. The parallel coordinate plot in action: Design and use for geographic visualization. Computational Statistics and Data Analysis. 2003;43(4):605-619.
7. Siirtola H. Combining parallel coordinates with the reorderable matrix. In Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization. July 2003;63-74.
8. Holten D, van Wijk JJ. Evaluation of cluster identification performance for different PCP variants. Computer Graphics Forum. 2010;29(3):793-802.
9. Harter JM, Wu X, Alabi OS, Phadke M, Pinto L, Dougherty D, Petersen H, Bass S, Taylor II RM. Increasing the perceptual salience of relationships in parallel coordinate plots. In Proceedings of SPIE Visualization and Data Analysis 2012. January 2012.
10. Krzywinksi M, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Research. September 2009;19(9):1639-1645.
11. Kosara R, Bendix F, Hauser H. Parallel sets: Interactive exploration and visual analysis of categorical data. IEEE Transactions on Visualization and Computer Graphics. July/August 2006;12(4):558-568.
12. Schonlau M. Visualizing categorical data arising in the health sciences using hammock plots. In Proceedings of the Section on Statistical Graphics, American Statistical Association. 2003.
13. Fua YH, Ward MRE. Hierarchical parallel coordinates for exploration of large datasets. In Proceedings of the Conference on Visualization '99: Celebrating Ten Years. 1999;43-50.
14. Heinrich J, Weiskopf D. Continuous parallel coordinates. IEEE Transactions on Visualization and Computer Graphics. 2009;15(6):1531-1538.
15. Zhou H, Yuan X, Qu H, Cui W, Chen B. Visual clustering in parallel coordinates. Computer Graphics Forum. May 2008;27(3):1047-1054.
16. Heinrich J, Luo Y, Kirkpatrick AE, Zhange H, Weiskopf D. Evaluation of a bundling technique for parallel coordinates. In Proceedings of the International Conference on Computer Graphics Theory and Applications and International Conference on Information Visualization Theory and Applications. 2012;594–602.
17. Ankerst M, Berchtold S, Keim DA. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In Proceedings of the IEEE Symposium on Information Visualization. 1998;52-60.
18. Peng W, Ward MO, Rundensteiner EA. Clutter reduction in multi-dimensional data visualization using dimension reordering. In Proceedings of the IEEE Symposium on Information Visualization. 2004;89-96.
19. Seo J, Shneiderman B. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In Proceedings of the IEEE Symposium on Information Visualization. 2004;65-72.
20. Bostock M, Ogievetsky V, Heer J. D3: Data-driven documents. IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis). 2011;17(12).
21. Tufte ER. The Visual Display of Quantitative Information. 2nd ed. Cheshire, CN:Graphics Press. 2001.
22. Office for National Statistics. 2011 Census: Ethnic group, local authorities in England and Wales. 2012.
23. Sinha R, Anderson DE, McDonals SS, Greenwald P. Cancer risk and diet in India. Journal of Postgraduate Medicine. July-September 2003;49(3).

POSTER PRESENTATIONS

Visual Analysis of Infection Surveillance

Penny B. Cooper[a], Nora M. Haney[b], José M. Morey[c]

[a] Augusta Health; Fishersville, VA
[b]Tulane University School of Medicine; New Orleans, LA
[c]Department of Radiology and Imaging, University of Virginia; Charlottesville, VA

Corresponding Author: Penny B. Cooper
pcooper@augustahealth.com
(540) 932-5564

Keywords: Infection Prevention Programs, Surveillance, Visual Analytics

MANUSCRIPT

Background and Significance

Hospital acquired infections are of particular concern for infection preventionist (IP) professionals. Surveillance is a critical part of infection control. [1-3] Increased electronic health record (EHR) adoption provides the technology setting for improving surveillance efforts through automation. [1,2,4,5] Therefore, Electronic Surveillance Systems (ESSs) capable of tracking trends of infections have been demonstrated to have moderate to excellent utility when compared to non-automated surveillance systems. [2,3,6]

Infection preventionist (IP) professionals spend a majority of their time ranging from 39-45% dedicated to infection surveillance and analysis. [7,8] A study by Grota et al. found that all IPs, regardless of ESS status at their institution, did not display differences in the way they spent their time concerning infection prevention activity or location. [3] These studies may collaboratively indicate that traditional methods featured in most surveillance systems are not effective at decreasing time spent on monitoring infection.

There is limited evidence of implemented visualization tools customized for a community hospital to aid IPs in the detection and remediation of infectious disease. This case study is an example of how visual analytics in healthcare can impact infection prevention and control efforts. The model was developed at Augusta Health, a 255-bed community hospital staffed by approximately 180 physicians and 2,300 employees. The hospital has 12,000 admissions annually, totaling 52,000 inpatient days, and has 60,000 annual emergency department encounters. In addition to acute and surgical care, the hospital also offers psychiatric, skilled nursing, and rehabilitative services.

Objective

Our objective was to create an in-house visual representation of infectious disease data analytics with the goals of decreasing data mining and analysis by IPs, as well as distribute infection data to physicians, nurses, and ancillary staff in a tangible mechanism.

Materials and Methods

The data visualization aspect is a feature set of a custom infection surveillance application under development at Augusta Health. Data is updated on a regular basis from the MEDITECH data repository relational database system that serves as the platform for data collection and analysis for the tool. The data repository serves as a long-term archive of all EMR data. The transfer of data from the proprietary EMR to the data repository is near real time. Because the hospital's market share is approximately 70%, the majority of admitted patients have administrative and clinical data from previous visits assisting in the collection of historical patient data for review.

The visualization combines the data elements of positive organisms and their attribution to specific patients and their occupied rooms. The count of infected patient stays in a room is superimposed over the hospital floor plans showing the pattern and spread of infections.

Results

The resulting interface is shown in Figure 1. The intensity of the color was designed to be proportional to the number of positive patients that have occupied the room within a specific time period. The date defaults to a range including the last 365 days but can be modified as desired. Furthermore, the visualization can also be filtered for specific organism, such as Methicillin- Resistant S Aureus or Clostridium difficile as shown in Figure 2.

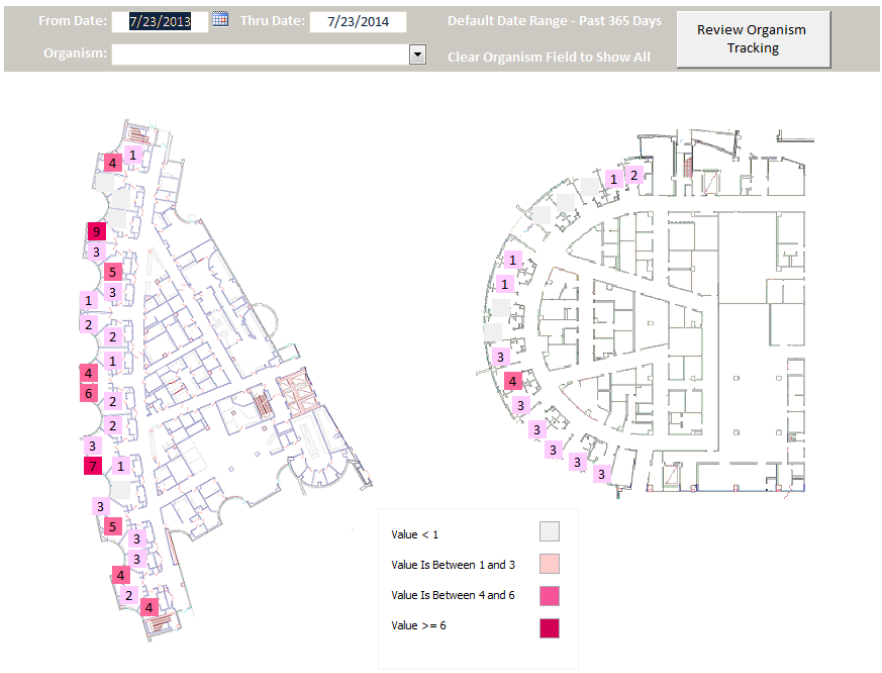


Figure 1: Visualization of the number of patients infected in each hospital room based on the last 365 days

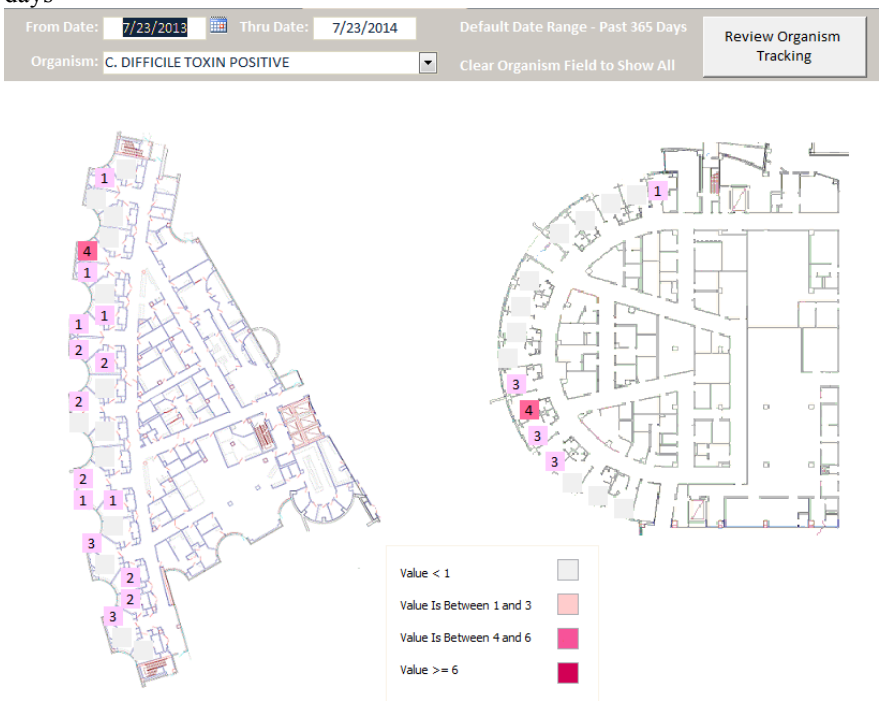


Figure 2: Visualization of the number of patients infected in each hospital room based on the last 365 days for C. Difficile Toxin

Discussion

The role of the IP professional is expanding from just surveillance of infection and provider behavior to include implementation of processes aimed at decreasing infection rates. [7,9] Therefore, instantaneous graphical representation of surveillance data may benefit IPs by easing their role of data mining and

analysis and allow more time for interventions. The visualization can assist the IP to see where preventative measures have failed and/or if environmental conditions need addressing. Graphing over time also assists IPs in detecting patterns. Like the room visualization, this can also be modified to show a custom date range and or filtered for specific organism. Grota et al. described ESSs as used frequently by IPs for data mining, but minimally by other infection prevention staff. [3] The graphical surveillance is automated in near real time. This feature allows IP professionals to relay important infection related information to physicians, nurses, and ancillary staff without further translation of data into decipherable information.

A significant strength of this custom application is that it can be implemented using off the shelf products populated with readily available EMR data. Limitations include the necessity to fit the model to a specific hospital floor plan however once the concept was developed and floor plans retrieved, this aspect was relatively quick work.

A future expansion of this application is planned to include the capability to visually track the patient to multiple rooms. An example of this future expansion is shown in Figure 2. This feature is necessary to ensure environmental measures have been executed after an infectious patient has occupied a room. This feature would be even more critical when an infection is discovered after the patient had been hospitalized for several days and had been transferred to multiple units during the period.

Conclusion

Research indicates that sophisticated systems have the potential to improve surveillance efforts while helping to control costs. [3] Improvements in Health IT infrastructure have set the stage to take advantage of improved surveillance through visualizations. Custom solutions integrated with the hospital EHR at Augusta Health have been developed and are currently in place including a tool used to estimate a patient's predisposition to clostridium difficile infection. [10] Feedback from this development indicates that additional customizations will be supported.

Conflict of Interest

The authors have no conflict of interest to disclose.

References

1. Hota B, Jones RC, Schwartz DN. Informatics and infectious diseases: What is the connection and efficacy of information technology tools for therapy and health care epidemiology? *Am.J.Infect.Control*. 2008;36(3):S47-S56.

2. Wright MO, Perencevich EN, Novak C, Hebden JN, Standiford HC, Harris AD. Preliminary assessment of an automated surveillance system for infection control. *Infect Control Hosp Epidemiol*. 2004;25(4):325-332.

3. Grota PG, Stone PW, Jordan S, Pogorzelska M, Larson E. Electronic surveillance systems in infection prevention: Organizational support, program characteristics, and user satisfaction. *Am J Infect Control*. 2010;38(7):509-514.

4. Atreja A, Gordon SM, Pollock DA, Olmsted RN, Brennan PJ. Opportunities and challenges in utilizing electronic health records for infection surveillance, prevention, and control. *Am J Infect Control*. 2008;36(3):S37-S46.

5. Leal J, Laupland KB. Validity of electronic surveillance systems: A systematic review. *J Hosp Infect*. 2008;69(3):220-229.

6. Halpin H, Shortell SM, Milstein A, Vanneman M. Hospital adoption of automated surveillance technology and the implementation of infection prevention and control programs. *Am J Infect Control*. 2011;39(4):270-276.

7. Stone PW, Dick A, Pogorzelska M, Horan TC, Furuya EY, Larson E. Staffing and structure of infection prevention and control programs. *Am J Infect Control*. 2009;37(5):351-357.

8. O'Boyle C, Jackson M, Henly SJ. Staffing requirements for infection control programs in US health care facilities: Delphi project. *Am J Infect Control*. 2002;30(6):321-333.

9. Curchoe R, Fabrey L, LeBlanc M. The changing role of infection prevention practice as documented by the certification board of infection control and epidemiology practice analysis survey. *Am J Infect Control*. 2008;36(4):241-249.

10. Cooper PB, Heuer AJ, Warren CA. Electronic screening of patients for predisposition to clostridium difficile infection in a community hospital. *Am J Infect Control*. 2013;41(3):232-235.

Iterative Methodology and Usability Improvement for Clinical Decision Support

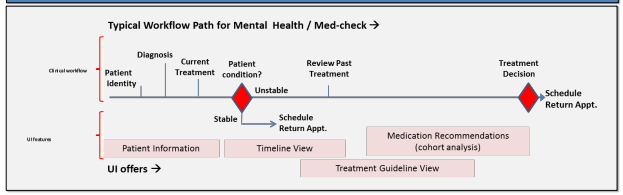
Fei Yu[1], Ketan K. Mane [2], Vincent Carrasco[1], and Javed Mostafa [1]

[1]Laboratory of Advanced Informatics Research, University of North Carolina and Renaissance Computing Institute, Chapel Hill, NC [2]

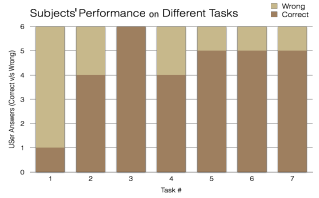
Summary

A visual-based interactive clinical decision support (CDS) tool was designed and is currently undergoing phased usability testing. Simultaneously, an iterative methodology for evaluating the usability of CDS user-interfaces (UI) is being developed and refined. Phase I of the project developed and evaluated a UI. Results revealed methodological challenges and the need to better align the UI with clinical workflows. This poster covers lessons learned regarding the specific gaps in the evaluation methodology and some UI refinements. The current UI is called Phase II UI.

Clinical Workflow



Phase I Evaluation Results



Task#	Data view needed complete task
1	Timeline
2	Patient Rx Profile
3	Rx Recommendation
4	Treatment Guideline
5	Treatment Guideline
6	Filtering
7	Predictive

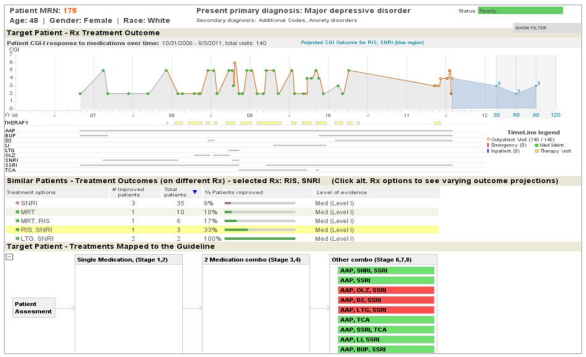
Data View	Efficiency in correctly communicating the info	Like-ability Factor
Timeline	17% (1/6)	★★★★★
Patient Rx Profile	68%	★★★★
Rx Recommendation	100%	★★★
Treatment Guideline	68% - 85 %	★★
Filtering	85%	★★½
Predictive Insight	85%	★★★★★

Medicine is Risk-Averse
- Favors knowing more

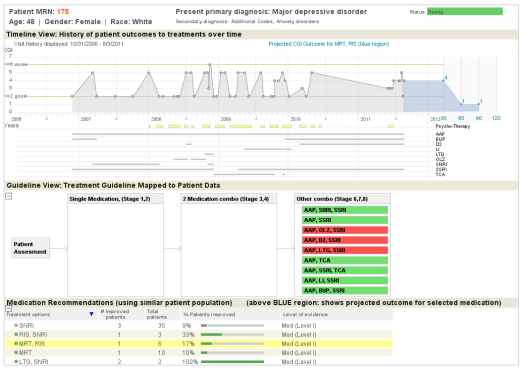
Challenge to UI design
- One size doesn't fit all tasks

UI Comparison

Phase I



Phase II



Methodology Aims and Steps

- Improve clinical workflow analysis and documentation
- Redesign usability tools (pre & post test questionnaires, task questions, orientation, etc.)

Refined Methodology

Quantify user experience

For each user task, a baseline was established:
- **Time** needed to complete the task
- **Steps** needed to complete the task
- **Defined** successful task completion
- **Documented** user strategies for task completion

Measure QUALITY user experience

- Redesigned Pre-test and Post-test questionnaires to reflect UI refinements

Future Steps: Phase II

- **Apply** the new evaluation methodology with a expanded pool of subjects and new tasks
- **Analyze** findings and generate a research report
- **Refine** UI

Reference:
Mane, K., Carrasco, V., & Mostafa, J. (2013). Visual-based Interactive Clinical Decision Support Tools: Developing User-centric Approaches , In *Proceedings of the AMIA Visual Analytics Workshop*, Nov., 16, Washington DC.

Acknowledgments: This project was supported by funds from UNC-CTSA, RENCI, the Agency for Healthcare Research and Quality (AHRQ) Grant R21HS019023 and NSF Award # 1117286.

2014 Workshop on Visual Analytics in Healthcare

A Comparative Analysis of Clinical Data Visualization Development Strategies: Build vs. Buy.

Jaehoon Lee, PhD[1], Thomas A. Oniki, PhD[1], John R. Holmen, PhD[1,2], Peter J. Haug, MD[1,2], Stanley M. Huff, MD[1,2]
[1]Intermountain Healthcare, Murray, UT; [2]University of Utah, Salt Lake City, UT

Abstract

In this paper we describe a comparative analysis of two clinical data visualization development strategies: 1) use of an existing visualization tool and 2) development of new open source software (S/W). We evaluated two visualization projects at Intermountain Healthcare, which used the analytic health repository (AHR) as a common data source but adopted different visualization tools: Tableau and d3.js. We analyzed pros and cons of the two approaches in terms of clinical knowledge representation, data management, and development productivity.

Background

In developing clinical data visualization, two general approaches have been widely used. One is to develop clinical data visualizations using an existing open source or commercial tool (e.g. Tableau [1]) that provides a set of tool-kits that cover most necessary functionalities, including data retrieval and processing, front-end user interfaces, and visualizations [2]. The advantage of this approach is high productivity and a systematic development process, although functionality may be limited by the toolkits. The other approach is to develop a homegrown visualization application using open source S/W libraries (e.g. d3.js [3]). This provides more flexibility but requires additional S/W development efforts. Although there have been several implementations based on these approaches, to the best of our knowledge there has been no study to investigate pros and cons of the two approaches.

Method

We conducted a comparative analysis of the approaches based on two visualization projects at Intermountain Healthcare. The two projects used the AHR as a common data source, which is a distilled copy of clinical data from our enterprise data warehouse designed to simplify population-based research activities. Its tables contain essential clinical data elements such as patient, encounter, vital sign, laboratory result, diagnosis, procedure, and medication. Since each table represents a very large data set, we created materialized views to prepopulate data as aggregated measures for fast responses to user inputs. Based on the materialized views, we developed two visualizations: 1) an open source based web application using d3.js [5], and 2) a Tableau based dashboard, which is a popular and richly functioned commercial visualization tool (See Table 1). Both applications are running within Intermountain's firewall (See Figure 1). We analyzed pros and cons of the approaches (See Table 2).

Table 1. Aggregated data types in the AHR and implemented visualizations

Category	Data type	Visualization type
Patient demographics	#Patient	*Line chart, bar chart; +Parallel set, donut chart
Encounter	#Encounter, #Patient	*Line chart, bar chart, bubble chart
Vitals / Laboratory result	#Record, #Patient, Descriptive statistics, Bin/Frequency	*Line chart, bar chart, histogram, bubble chart; +Box plot, chord diagram
Diagnosis / Procedure / Medication	#Record, #Patient	*Line chart, bar chart, bubble chart; +Word cloud, tree map

*Tableau and d3.js, +D3.js

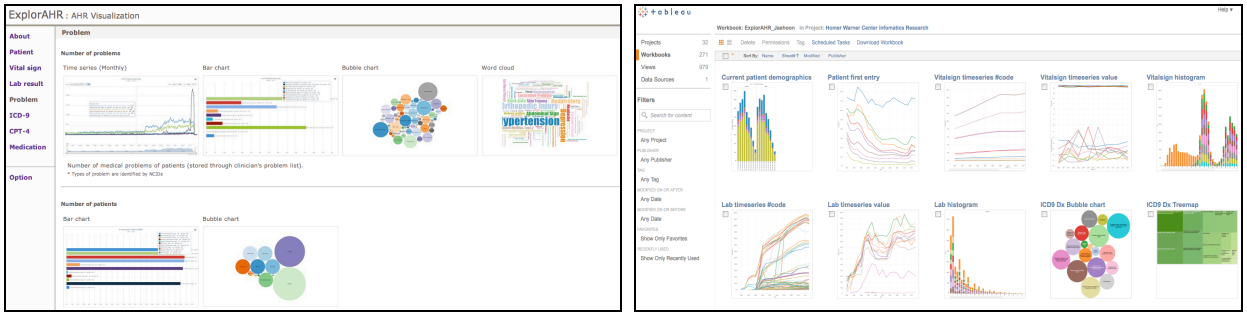


Figure 1. a) d3.js based visualization; b) Tableau dashboard

Table 2. Differences of implementation between d3.js and Tableau

Category	d3.js based development	Tableau based development
Visualization generation	Selected visualizations from more than 219 visualization types in github.com and downloaded source codes. Additional data preprocessing and data format validation were done to feed AHR data to the libraries.	Selected visualizations from Tableau provided 24 visualization types. Automated generation of visualizations and data format validation was supported by Tableau authoring environment.
User interface	Developed customized web pages and filters using HTML5[#]	Used Tableau dashboards and filters
Data storage	JSON[*] file	Tableau data extract[+]
Data retrieval and processing	A Java program was developed to retrieve data from the materialized views and generate JSON files.	Imported data from the materialized views using Tableau data connection. Once imported, Tableau maintains the data as Tableau data extract at both development and operation phase.
Knowledge management	Terminologies and class of hierarchies were hard coded.	Mapped clinical concepts to surface forms (general names) using alias function, and create class of hierarchies using group function. These were done manually when creating filters at development phase.
Development productivity	An entire S/W development process from design to testing as well as a web developer and a database analyst were required.	A general Tableau user with basic database knowledge could develop visualizations. A Tableau server and administrator were needed. No issue with S/W engineering and process.

[*]JavaScript Object Notation, [+]A specialized data format for internal use in Tableau, [#]Hypertext Markup Language fifth revision

Conclusion and Discussion

Pros and cons: Overall, the advantage of using Tableau is high productivity at the development and operation phases, whereas the open source development strategy provides the flexibility of using a variety of visualizations and an extensible S/W architecture. The Tableau based approach required a relatively shorter development time. Cost is not a directly comparable factor since Tableau requires the purchase of commercial license while d3.js is free from a license cost but requires more S/W engineering effort and resources. For both implementations, since they commonly used prepopulated data in the materialize views, performance (response time to user inputs) was acceptable.

Integrated clinical knowledge support is required. Since both Tableau and d3.js are general-purpose visualization tools and do not deal with domain knowledge, a knowledge base may be needed to support clinical knowledge such as standard terminologies and clinical data models. Class hierarchies are important for underlying concepts to support filters and zoom in/out, which are essential features of data visualization. Semantic linkages between concepts may be useful when a user changes visualization perspectives (e.g. moves from cardiovascular diseases to cardiovascular system procedures). To realize these functionalities in d3.js based applications, web development is required to integrate user interfaces with the knowledge base. In the case of Tableau, although the Tableau dashboard does not support direct connection with an external knowledge base, integration is feasible by utilizing Tableau's Javascript APIs (Application Programming Interfaces), which enables a web application to embed and control remote Tableau visualization objects in HTML.

A hybrid development strategy may be useful. In the case of implementing multiple visualizations, a hybrid approach could take the advantages from the both approaches. Developers may use Tableau for basic visualizations and use d3.js for specific visualizations which are not supported by Tableau. A clinical knowledge base and prepopulated data infrastructure will add value by providing a standardized and reusable way of managing clinical knowledge and data in heterogeneous visualization tools.

References

1. http://www.tableausoftware.com/
2. Wohlfart E, Aigner W, Bertone A, Miksch S: Comparing information visualization tools focusing on the temporal dimensions, information visualisation, 2008. IV '08. 12th International Conference. 2008: 69-74.
3. Bostock M, Ogievetsky V, Heer J: D3: Data-driven documents, IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis). 2011, 17(12): 2301-2309.
4. Lee J, Holmen JR, Catmull SL, Pollock SE, Haug PJ, Huff SM: Research oriented clinical data visualization based on the analytic health repository: A case study at Intermountain Healthcare, VIS 2013, Atlanta, GA.

Text Mining and Visualization to Explore E-Cigarette and Hookah-Related Social Media

Annie T. Chen
School of Information and Library Science
University of North Carolina
Chapel Hill, United States
atchen@email.unc.edu

Shu-Hong Zhu, Mike Conway
Department of Family and Preventive Medicine
University of California
San Diego, United States
szhu@ucsd.edu, mconway@ucsd.edu

***Abstract*—In recent years, e-cigarettes and hookah have risen in popularity. We take a novel approach to understanding the use and appeal of these tobacco-related products by applying text mining and visualization techniques to social media sources, comparing between the sources and making inferences about user experience. We describe two visualizations: an interactive heat map and a symptom polarity visualization. The heat map may be used to explore differences across forums in terms of contextual factors of health behavior: e-cigarette and hookah-related conversation, symptoms, quitting experience, health perceptions, sociality, sensory experience, setting, and time. The symptom polarity chart may be used to investigate the nature of symptoms that are reported.**

Keywords—text mining; visualization; heat map; social media; electronic cigarette (e-cigarette); hookah

I. Introduction

In recent years, researchers have begun to realize the value of social media as a data source to understand health-related phenomena, for purposes such as influenza surveillance and identification of adverse effects from medications (e.g. [1-2]). This paper employs text mining and visualization techniques to compare social media discussions regarding two tobacco-related products: e-cigarettes and hookah. A number of studies have examined e-cigarette and hookah in social media. Hua et al. [3] identified symptoms reported in discussion forums, while Myslín et al. [4] used Twitter to analyse smoking behavior and public perceptions of e-cigarettes and hookah.

The growth in popularity of e-cigarettes has been documented in several developed countries, and is a growing focus of public health concern, especially in the United States and the European Union [5-6]. There has also been increasing concern about the use of hookah, a centuries old practice that is currently increasing in prevalence in the Middle East and worldwide [7-8]. The device is also called other names such as waterpipe, shisha, and hubble-bubble arghileh.

We are developing a system that integrates text mining and visualization techniques to support exploration of discussion content about e-cigarettes and hookah. In this paper, we describe two of the visualizations supported by the system.

II. Datasets

We employ content from three websites: Vapor|Talk, Hookah Forum, and Reddit. Vapor|Talk and Hookah Forum are popular online communities that are dedicated to e-cigarettes and hookah, respectively. Reddit is a generic platform that features subreddits focused on a broad spectrum of topics. We expected that these samples might differ on a variety of characteristics.

III. Health Behavior Context Heat Map

Previous comparisons of online communities have examined characteristics such as size, response rate, and typology [9]. We focus on a different type of characteristic: aspects of health behavior. Estimating differences in prevalence of aspects of health behavior across datasets can help match research questions to appropriate data sources.

To examine differences across datasets, we employed a heat map visualization. Heat maps are often used in genetics to display gene expression patterns (e.g. [10]). In a classic cluster heat map, one axis might represent samples, and the other, genes [11]. Cells are colored based on the level of expression of the gene in the corresponding sample.

The health behavior context heat map was based on this work. We select sub-forums on the websites which we compare along dimensions of interest: e-cigarette terminology, hookah terminology, symptoms, quitting experience, perceptions, health care practitioners, sociality, sensory experience, setting, time and cost. These dimensions represent contextual aspects of health behavior. Differences in color intensity provide insight concerning interests and/or behavior patterns of the samples. Each of the rows corresponds to one dimension, and the columns, to one dataset (Fig. 1).

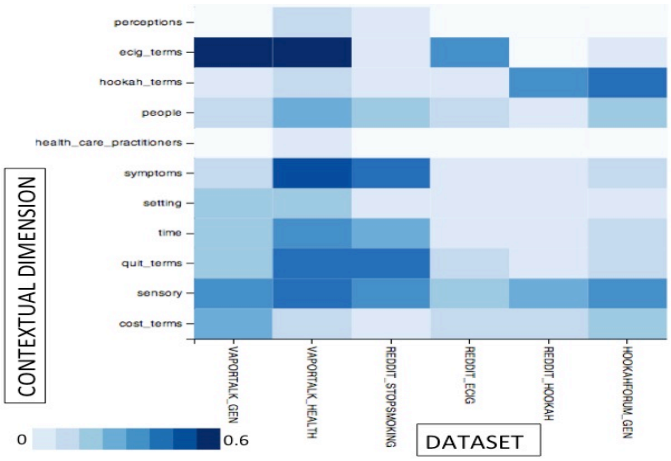


Fig. 1. Health Behavior Context Heat Map

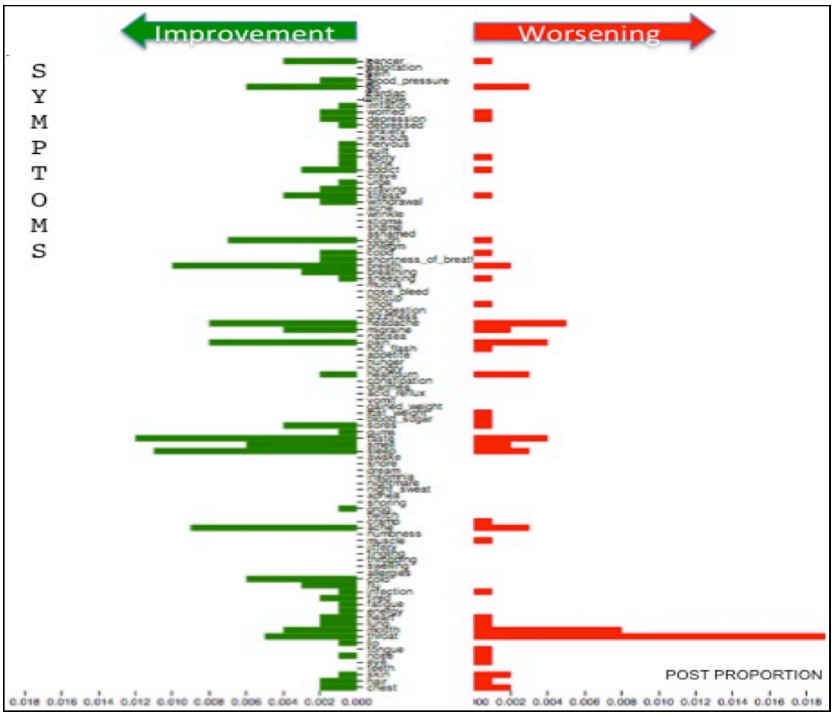


Fig. 2 Symptom Polarity Chart: Vapor|Talk Health & Safety.

As expected, e-cigarette related content is most popular in the Vapor|Talk forums and on the electronic_cigarette subreddit, and hookah content is popular in the hookah forums. Quitting and symptoms are mentioned most often in the Vapor|Talk Health & Safety and the Stopsmoking subreddit. The cells of highest density for people, time and health perceptions also come from these two samples.

IV. SYMPTOM POLARITY CHART

The second visualization is a provisional symptom polarity chart based on a subset of forum data. A rule-based extraction system was used to identify symptoms and their polarities. A report was considered positive if a positive change in health status was reported, such as a decrease in frequency of coughing; and negative if a negative change, such as a sore throat, was reported. Symptom identification is based on a knowledge base developed for our system, which maps symptoms and their synonyms, to organ systems. Clicking on the individual bars results in the posts containing the symptom reports being displayed. One of the sub-forums with the highest density cells in the heat map, Vapor|Talk Health & Safety, is depicted in Fig. 2. Members discuss a diverse array of physiological experiences, including headaches, sore throat, and changes in their experience of taste.

V. CONCLUSION

In this paper, we explore how text mining and visualization techniques might be integrated to facilitate comparison across social media datasets. Heatmaps highlight differences across samples and inform dataset selection, and visualization of symptom polarity mentions may be used to make clinically relevant inferences.

This work was sponsored by a grant from the National Cancer Institute U01 CA154280.

REFERENCES

[1] C. D. Corley, D. J. Cook, A. R. Mikler, and K. P. Singh. Text and structural data mining of influenza mentions in Web and social media. Int J Environ Res Publ Health, vol. 7, no. 2, pp. 596–615, February, 2010.

[2] A. T. Chen. Patient experience in online support forums: Modeling interpersonal interactions and medication use. ACL-SRW 2013.

[3] M. Hua, M. Alfi, and P. Talbot. Health-related effects reported by electronic cigarette users in online forums. J Med Internet Res, vol. 15, no. 4, e59, April, 2013.

[4] M. Myslín, S.-H. Zhu, W. Chapman, and M. Conway. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res, vol. 15, no. 8, e174, August, 2013.

[5] S.-H. Zhu, A. Gamst, M. Lee, S. Cummins, L. Yin, and L. Zoref. The use and perception of electronic cigarettes and snus among the U. S. population. PLoS ONE, vol. 8, no. 10, e79332, October, 2013.

[6] C. I. Vardavas, F. T. Filippidis, and I. T. Agaku. Determinants and prevalence of e-cigarette use throughout the European Union: a secondary analysis of 26 566 youth and adults from 27 Countries. Tob Control, in press.

[7] W. Maziak, K. Ward, S. Afifi, and T. Eissenberg, Tobacco smoking using a waterpipe: a re-emerging strain in a global epidemic. Tob Control, vol. 13, no. 4, pp. 327–333, December, 2004.

[8] M. P. Martinasek, R. J. McDermott, and I. Martini. Waterpipe (hookah) tobacco smoking among youth. Curr Probl Pediat Adol Health Care, vol. 41, no. 2, pp. 34–57, February, 2011.

[9] K. T. Durant, A. T. McCray, and C. Safran. Identifying temporal changes and topics that promote growth within online communities: A study of six online cancer forums. Int J Comput Models Algorithms Med, vol. 2, no. 2, pp. 1-22, April, 2011.

[10] D. Merico, R. Isserlin, R. Stueker, A. Emili, and G. D. Bader. Enrichment map: A network-based method for gene-set enrichment visualization and interpretation. PLoS ONE, vol. 5, no. 11, e13984, November, 2010.

[11] L. Wilkinson, and M. Friendly. The history of the cluster heat map. Amer Statist, vol. 63, no. 2, pp. 179-184, May, 2008.

Interactive Demographic Visualization of Multiple Facilities across Time

Hyunggu Jung, MS[1], E. Sally Lee, PhD[2], Hossein Estiri, PhD[2], Kari A. Stephens, PhD[1,2]
[1]Department of Biomedical Informatics and Medical Education, [2]Institute of Translational Heath Sciences, University of Washington, Seattle, WA

Abstract

We present an interactive demographic visualization of multiple facilities across time that can be generalized. It provides not only users with an overview of patient demographic information of multiple facilities across time, but also offers detailed breakdowns of patient demographics such as age, gender, language, and race of a selected facility by allowing users to choose their view preference. When users mouse over a specific area of the bar graph of summary statistics on the facilities, four charts corresponding to the breakdown of demographic information (i.e., age, gender, language, and race) are updated dynamically. Furthermore, the visualization displays summary statistics of the selected facility that users are viewing.

Introduction

The electronic health record (EHR) data are complicated, reflecting a variety of patient health information, such as patient demographics, progress notes, medications, vital signs, laboratory data, etc. Studies have evaluated visual analytic tools for navigating and analyzing such health data, targeting ease of use of these complicated datasets. In one study, Wang and colleagues focused on navigating multiple records of categorical temporal data from the EHR [1]. The prototype of their visualization tool enabled users to align, rank, and filter the results of queries [1]. In another study, Zhang and colleagues focused on analyzing patient cohort data, building an interactive visualization application that enabled clinicians to explore patient cohort data by visualizing and refining cohorts [2].

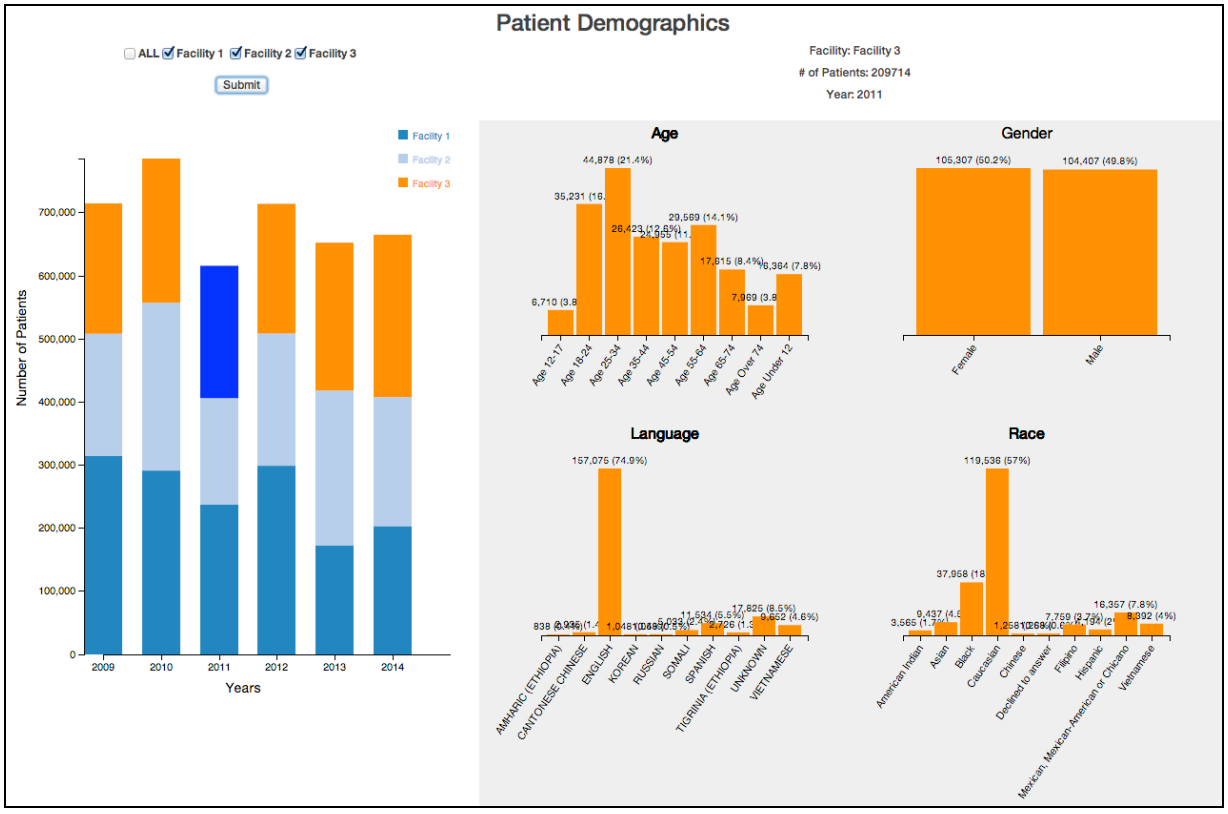


Figure 1. Interactive demographic visualization with four components: a set of check boxes for allowing users to determine any combinations of facilities based on their interest, a dashboard displaying the summary of the selected data with the name of the facility, the number of patients, and the year of the selected data, and four charts representing the patient demographic information (i.e. age, gender, language, and race).

In this paper, we introduce an interactive demographic visualization of multiple facilities across time with temporal patient demographics data to support users whose aim is to explore the existing clinical data to later conduct research. Based on the golden rule of data visualization such as overview, zoom and filter, and detail-on-demand [3, 4], we applied the technique of "Level of Detail (LOD)" in creating our visualization. For LOD, in addition to the summary view of a cohort of the specific population [2], our visualization provides users with an additional four charts to represent detailed information of patient demographics by age, gender, language, and race.

Target Users

Our visualization is targeted to individuals who use data for generated through clinical care for research. The potential target users can be any biomedical researcher, such as faculty, residents, fellows, graduate students, or research staff. For such users, it is crucial to explore the information available about the data in order to generate questions, test feasibility, and conduct biomedical research on specific populations. For example, some clinical studies have age restrictions, others have gender restrictions, and some require information about the spoken language as they can only recruit English speakers.

Visualization

We generated visualizations with D3 Javascript library (http://d3js.org/). The visualization demonstrates patient demographic information of multiple facilities across time as illustrated in Figure 1. Users can choose any combination of facilities to explore the overview of patient demographic information. When users mouse over on the specific area of the bar graph, four charts corresponding to breakdowns of demographic information (i.e. age, gender, language, and race) are updated dynamically. For consistency, the four charts were colored with the same color of the selected area of the bar graph. On the top of each bar on the charts, the number of patients and the percentage of all the number of patients are displayed.

Conclusion

We have presented an interactive demographic visualization of EHR data from multiple facilities across time. Our visualization provided users with two main features to express "Level of Detail": 1) allowing users to choose their view preference by filtering the source of data, and 2) displaying the detailed information based on their interest by hovering a mouse on the bar chart. These features allow users to navigate patient demographics such as age, gender, language, and race via four charts updated dynamically. While this prototype offers users dynamic access to basic demographic EHR data, validating its features would confirm potential for dissemination of use. This prototype provides a useful starting point to evaluate and iterate visualizations for broader dissemination [5, 6]. The dynamic nature of the web-based visualization method provides promise for researchers to interact with complex EHR data efficiently and easily and speed use of EHR data for discovery.

References

1. Wang, T. D., Plaisant, C., Quinn, A. J., Stanchak, R., Murphy, S., & Shneiderman, B. (2008, April). Aligning temporal data by sentinel events: discovering patterns in electronic health records. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 457-466). ACM.
2. Zhang, Z., Gotz, D., & Perer, A. (2012, October). Interactive visual patient cohort analysis. In *Proc. IEEE Visual Analytics in Health Care (VAHC) Workshop*.
3. An, J., Wu, Z., Chen, H., Lu, X., & Duan, H. (2010, October). Level of detail navigation and visualization of electronic health records. In Biomedical Engineering and Informatics (BMEI), 2010 3rd International Conference on (Vol. 6, pp. 2516-2519). IEEE.
4. Wang, T. D., Wongsuphasawat, K., Plaisant, C., & Shneiderman, B. (2010, November). Visual information seeking in multiple electronic health records: design recommendations and a process model. In Proceedings of the 1st ACM International Health Informatics Symposium (pp. 46-55). ACM.
5. Fonteyn ME, Kuipers B, Grobe SJ. A Description of Think Aloud Method and Protocol Analysis. Qual. Health Res 1993; 3(4): 430–441.
6. J. Nielsen. Heuristic evaluation. In J. Nielsen, and R.L. Mack, (Eds.). Usability Inspection Methods. John Wiley. 1994-& Sons, NY, USA, 25-61.

Envisaging Biomedicine: Visualization of the Clinical and Scholarly Ecosystem using Bibliographic Metadata

Terrie R. Wheeler, MLS[1]; Karen Gutzman, MLS[2]; Michael Bales, PhD[1]; Paul Albert, MLS[1]; Kristi Holmes, PhD[2]

[1]Weill Cornell Medical College, New York, NY; [2]Galter Health Sciences Library, Feinberg School of Medicine, Northwestern University, Chicago, IL

Summary

This poster will describe some of the reports and visualizations that are possible using bibliographic data, resources needed for this work, the role of the library in helping individuals and groups accomplish this effort, and some strategies for visualizing data that go beyond basic publication data.

Introduction and Background

The ability to understand the real impact of research and clinical activities in the modern biomedical academic environment is a Sisyphean task, given the increasing complexity of today's research and clinical care landscape. The sheer amount of relevant, related data from commercial and open data aggregators, as well as from the enterprise itself in systems such as the EHR, data warehouse, and research information systems, has never been greater. The scholarly publishing process is obviously a very important step of disseminating scholarly discoveries to the broader community. This bibliographic metadata can also play a valuable role in helping an institution gain insights about its research and clinical activities that might not be discernable from other means, including some of the available institution-level data stores. These insights allow both investigators and the institutions themselves to convey the benefits and impact of their research and clinical efforts to stakeholders, thus providing valuable information that can be used for benchmarking, forecasting, and strategic planning activities.

Telling a Story

Bibliographic data can be mined and visualized using a variety of techniques to gain a better understanding of a variety of critical facets of research and clinical care. Research Information Systems can offer "out of the box" visualizations to yield a better understanding of the expertise and collaborations by an investigator, or among a group of investigators. Bibliographic metadata can be harvested from open or commercial data aggregators and analyzed to characterize geographic scholarly dissemination and knowledge transfer as well as more broad dissemination of health and research discoveries to the general public. Bibliometric metadata can also be used to accomplish detailed trend analysis locally as well as across all of science for benchmarking and strategic planning purposes. Visualizing the linkage between publications and grant data enables investigators and grant funding organizations to better comprehend the impact of their grant portfolios, and the efficacy of their funding decisions. Trending of these linkages over time may help identify promising areas for future research. Deeper investigation of this data could lend a better understanding of effective scholarly practice in a discipline.

There are a wide range of data sources and tools available to accomplish this work. These resources can vary widely, in terms of completeness, features, cost, and ease of use. The importance of good data cannot be overemphasized. Likewise, a variety of approaches can be employed to obtain quality reports and visualizations of research and clinical activities. Open source and commercial research information systems can offer features for visualizing various data contained in their systems, accomplishing social network analyses and contributing toward a smoother workflow, overall.

Moving Forward

Today's modern research and medical libraries provide resources and expertise to their campuses that can be leveraged to accomplish data analyses and visualizations. The role of the librarian in accomplishing and supporting this work is essential, especially given the critical importance of data integrity and the role that librarians already play as valued team members on significant enterprise-level information projects. While still in its infancy and often not broadly available, many libraries are committing resources and staff to support this work on behalf of their institution. Libraries offer the campus a perfect combination of expertise, perspective, and resources to help support and advise their assessment and visualization of research impact across the peer-reviewed literature and beyond.

LIVE DEMONSTRATIONS

Visualization and prediction of diabetes disease progression along with modifiable factors using Diabetes Complications Severity Index (DCSI)

Muhammad T. Dastagir, MD, Keith M. Glassford, PhD, Shalom Halevy, PhD, Erik R. Smith, Thomas J. Van Gilder, MD, JD, MPH
Certify Data Systems

Abstract

Diabetes mellitus not only continues to be a leading cause of morbidity and mortality in the United States, but is also a major contributor to the rising cost of health care. Poor understanding and inadequate communication of disease progression and modifiable factors from both physicians and patients is one of the major reasons for the suboptimal control of diabetes. Our goal is to help patients and physicians not only understand their current disease progression, but to also help them predict disease progression based on modifiable behavioral factors. This will help them focus their efforts on the behaviors that are helpful in gaining better control of diabetes and slowing the disease progression, which will help reduce morbidity and health care spending.

Introduction

According to a National Diabetes Statistics Report from 2014[1], 9.3% of the U.S. population has diabetes mellitus. In 2010, diabetes was the seventh leading cause of death in the U.S. and in 2012, costs related to diabetes were estimated to be around 245 billion.

One of the reasons diabetes continues to be among the leading causes of morbidity and mortality is poor disease control. According to a National Health and Nutrition Examination Survey (NHANES), diabetes is usually poorly controlled, and only 42.3% of surveyed patients had target HBA1c levels[2]. One of the barriers to adequate diabetes control is that patients have a poor understanding of the disease[3] and physicians are not helping their patients understand their health status[4]. Therefore, it is difficult for patients to determine the right behavior changes to make which results in poor motivation. Understanding the impact of change or lack of, [5] and the ability to self-monitor a disease's progression is likely to increase engagement[6]. Our goal was to show that claims or clinical data (or both) could be aggregated and visualized to show diabetes disease progression over time, using an accepted diabetes severity index.

Visualization of current disease course

In our model, we used the Diabetes Complications Severity Index (DCSI)[7] to measure and visualize disease progression. DCSI scores range from 0-13 based on the count and severity of diabetes complications. The DCSI score is presented longitudinally within each patient's medical profile.

Modifiable factors

Based on clinical literature review, we then identified modifiable factors with the greatest impact on disease progression and presented these in association with DCSI to help users understand the correlation. We utilized color-coding to demonstrate compliance or non-compliance for each modifiable factor in addition to a visual indication of recent changes made to these modifiable factors.

Predicting DCSI based on modifiable factors

We used linear regression to predict future DCSI based on modifiable and certain non-modifiable factors and presented the projected DCSI to the users (clinicians and patients). We also allowed users to view the impact of various behavioral changes (based on modifiable factors) to their future DCSI to help motivate behavior changes and slow the disease progression.

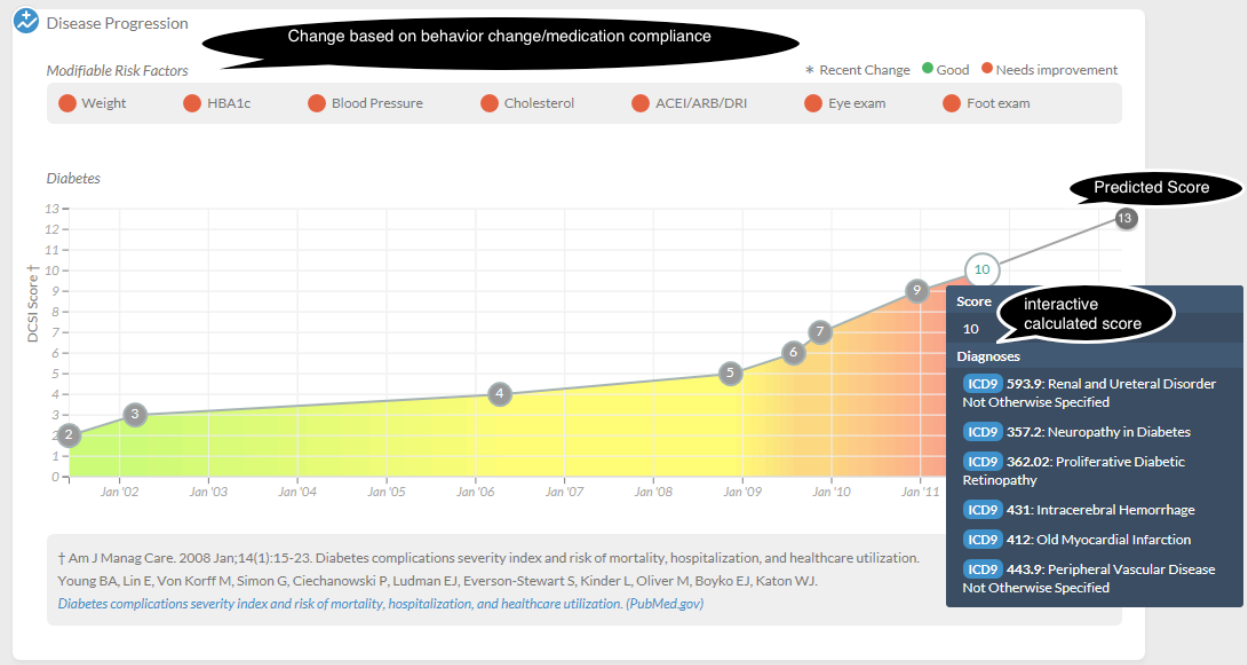


Figure 1. Visualization of disease progression and modifiable risk factors

Conclusion

We believe that presenting this data will not only improve a physician and patient's understanding of the disease but it will also emphasize the correlation between modifiable factors and disease progression. This enhanced understanding can be leveraged as a self-quantifiable, motivational tool, to initiate and sustain the behavior modifications necessary to slow disease progression. We have shown that claims and clinical data can be modeled and visualized to give a possible guide to physicians and their diabetic patients as they try to modify the course of diabetes. In future work, we will allow practicing physicians to use this data visualization and monitor behavior for signs of more efficient workflow and for measures of improved diabetes control both by both the physician and patient.

References

1. Centers for Disease Control and Prevention. National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States, 2014. Atlanta, GA: U.S. Department of Health and Human Services; 2014.
2. Saaddine JB, Cadwell B, Gregg EW *et al.* Improvements in diabetes processes of care and intermediate outcomes: United States, 1988–2002. Ann. Intern. Med. 144(7), 465–474 (2006).
3. Tessaro I, Smith SL, Rye S. Knowledge and perceptions of diabetes in an Appalachian population, Prev Chronic Dis. 2005 Apr; 2(2): A13.
4. Papp R, Borbas I, Dobos E, Bredehorst M, Jaruseviciene L, Vehko T, Balogh S. Perceptions of quality in primary health care: perspectives of patients and professionals based on focus group discussions, BMC Fam Pract. 2014 Jun 28; 15:128.
5. Janz NK, Becker MH. The Health Belief Model: A Decade Later, Health Educ Behav March 1984 vol. 11 no. 1 1-47.
6. Consolvo, S., Everitt, K.M., Smith, I., and Landay, J.A. Design requirements for technologies that encourage physical activity. Proceedings of CHI 2006, ACM Press (2006), 457-466.
7. Young BA, Lin E, Von Korff M *et al*. Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization, Am J Manag Care. 2008 Jan; 14(1): 15-23.

Weave: Utilization of InfoMaps and an Individual Record Tool for Patient Data Analysis

John Fallon[1]
University of Massachusetts Lowell
1 University Ave
Lowell, MA 01854

Georges Grinstein[2]
University of Massachusetts Lowell
1 University Ave
Lowell, MA 01854

***Abstract*— New tools have been developed in Weave to aid in the analysis of patient data. Weave is a web-based visualization system. The tools being introduced are InfoMaps and an Individual Record Tool. InfoMaps enables the visual querying of document collections that can be linked to other data. The Individual Record Tool is a prototype that provides a specialized view for patient data within the visualization environment Weave. The results are coordinated visualizations that permit looking at a single patient, related patients, and a complete collection of patients.**

Keywords—visualization; analysis; data; information visualization

I. INTRODUCTION

We have modified an existing visualization system, Weave, to allow for the visualization and analysis of both large patient data and developed a new visualization, called the Individual Record Tool (IRT) which supports the visualization of single patients or individuals. This new tool has been integrated with a novel tool in Weave called InfoMaps. Leveraging the two together along with already existing "standard" visualizations within Weave's sessioning system allows overviews of patient data while being able to drill down to single ones.

II. WEAVE

Weave is an open source web-based visualization platform designed to enable visualization of any available data by anyone for any purpose [1]. It provides coordinated visualizations, is targeted both for developers as well as end users, and supports integrated analysis and visualizations. Resulting visualizations can easily be disseminated in a web page. One of Weave's strengths is its underlying sessioning system. Every interaction a user makes with the system is recorded in a timeline of events occurring during the session from the time the user started Weave. This information can be recorded and used to replay analyses, can be used to tell a story, and can be emailed or shared with others. Utilizing Weave's built-in abilities, our additions support the analysis and dissemination of patient data.

III. INFOMAPS

InfoMaps is an information visualization tool designed for personal information management and for supporting data analysis [2]. We developed InfoMaps with the intention of linking large text document corpora with visualization and analysis. InfoMaps allows for the querying of various papers, books, journals, and any other text corpus that has been indexed. Some example views of InfoMaps querying a document cluster and then reporting back the results in various formats can be seen in Figure 1. The data returned from the InfoMaps tool can then be linked with the data being used in the visualizations in Weave. This allows for probing and selection to be linked across multiple tools that interact with InfoMaps. InfoMaps helps to bridge the barrier between text document collections and visualization and analysis of structured data.

Figure 1 InfoMaps Tool showing various query parameters and their results.

IV. INDIVIDUAL RECORD TOOL

The Individual Record Tool (IRT) specializes in viewing the details of one data point within the scope of many. It is based on Lifelines [3] and Tufte [4]. The tool's layout is quite flexible and can be programmed to have different looks and feel. An example of the tool being used within the scope of a Weave workspace can be seen in Figure 2. We can see distinguished upper area and lower area. The upper area is the profile area which provides room for a picture along with lines of text to be used as descriptions. The image and text can be loaded either from locally stored images on the server or a web accessible link, dynamically generated from columns of data, and can also contain customized user input. An analyst can lock the currently displayed record from changing as interactions take place in other Weave visualizations. The tool

1. john_fallon@student.uml.edu
2. grinstein@cs.uml.edu

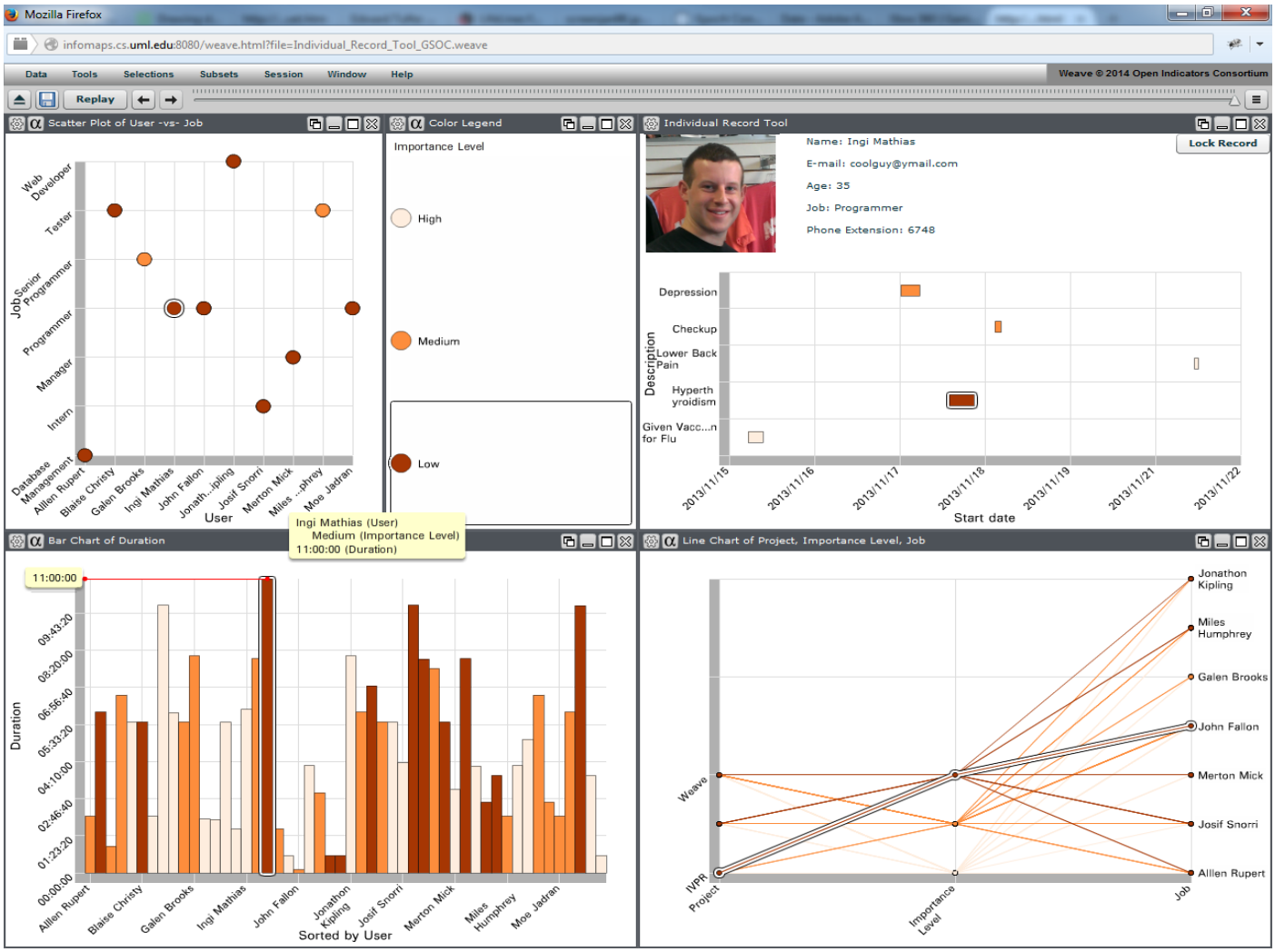


Figure 2 An example Weave workspace for examining patient data. The top right tool showcases the Individual Record Tool. It can be seen that there is linked probing between the tools. This also holds true for selection.

updates what record it is viewing as the user probes and selects single records within Weave.

The lower area of the tool contains the plotter which contains axes to indicate the passage of time and descriptions of the parameters of the record being viewed. Rectangles are drawn on the plotter based on the provided temporal data. A rectangle's length is indicative of when and for how long an event occurred as in Lifelines. The height and color of the plotted symbols can be customized by the user as well as adding other visual components. Linked probing and selection are supported. The IRT complements the rest of the tools in Weave's visualization tool suite to allow for in-depth analysis of patient data both at the global level (large data, maps, distributions, etc.) and at the local level by selecting individual patients and seeing them in the IRT.

A physician or clinician can look at an individual patient, look at similar patients in a scatterplot, select a few of these, see which documents relate to illnesses they have, read the documents, select one that discusses a protocol and see which selected patients it applies to. Weave thus goes from one to many to documents to subsets of the records back to documents and so on. Moving between text, data, and individual detailed views is fluidly provided.

ACKNOWLEDGMENT

We would like to thank the Google Summer of Code [5] for providing funding for the development of these additional tools in Weave.

REFERENCES

[1] Baumann, Alexander, and Georges Adviser-Grinstein. The design and implementation of weave: A session state driven, web-based visualization framework. University of Massachusetts Lowell, 2011.

[2] Kolman, Sebastin, et al. "InfoMaps: A Session Based Document Visualization and Analysis Tool." Information Visualisation (IV), 2012 16th International Conference on. IEEE, 2012.

[3] Plaisant, Catherine, et al. "LifeLines: visualizing personal histories." Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 1996.

[4] E. Tufte, S. M. Powsner, Graphical Summary of Patient Status, The Lancet 344 (8919) (1994) 386–389.

[5] "Google Summer of Code 2014 - Home Page." *Google Summer of Code 2014*. Google, n.d. Web. 03 Sept. 2014.

Demonstration of Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels

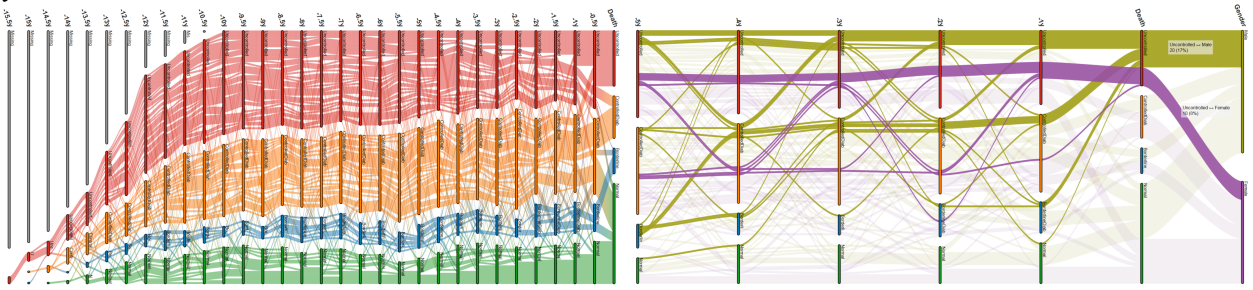
Hina Shah[1], David Borland[2], Eugenia McPeek Hinz[3], Vivian L. West[1], W. Ed Hammond[1]
[1]Duke Center for Health Informatics, Duke University, Durham, NC;
[2]RENCI, The University of North Carolina at Chapel Hill, Chapel Hill, NC;
[3]DHTS Duke Medicine, Durham, NC

Introduction

In this demonstration we present a visualization tool for a cohort of patients with diabetes (via ICD9 codes) from Duke University's data warehouse, visualizing their Hemoglobin A1c (HbA1c) levels over time, aligned by death, to explore trajectories of glycemic control. To the best of our knowledge, temporal visualization of glycemic control for a diabetic population standardized on death has not previously been presented. Our visualization groups HbA1c values into ordered categories of glycemic control, utilizing a method based on parallel sets and Sankey diagrams to view temporal patterns in HbA1c values. We incorporate a number of features for interactive data exploration like: viewing the progression of values either forwards or backwards in time, highlighting multiple subpopulations, coloring based on the category along each path in the data or at the beginning/end of each path, and the incorporation of demographic data, such as gender.

Methods

Data from Duke University's data warehouse were extracted using DEDUCE, an electronic health record (EHR) query tool developed at Duke University. The final cohort includes data from 121 patients with diabetes mellitus (with and without complications), a death indicator, prescribed antihyperglycemics, and at least 10 years of HbA1c laboratory values. We average HbA1c values over 6 month time intervals. In the case of missing HbA1c values within a 6 month period, we first attempt to impute a HbA1c value from average glucose (AG) values over that period of time if available, otherwise the previous HbA1c value (measured or imputed) is carried forward. HbA1c values are then categorized based on the severity of diabetes: Normal < 5.7, Borderline [5.7, 6.5), Controlled [6.5, 8), and Uncontrolled ≥ 8. The sampled data is time-aligned by the death event for each patient. The visual representation of diabetes progression propagates backwards in time initially. Time is represented as number of years before death in six month increments.



(a) (b)

Figure 1. Diabetes progression visualizations without and with a gender axis: (a) Overview visualization with paths colored by the current HbA1c at each time step, useful for emphasizing overall temporal trends. (b) Adding a gender axis and selecting two groups, we can compare the variability of males who were Uncontrolled at death (olive) to women who were Uncontrolled at death (purple). Men appear to have more variability over the 5-year period being visualized, as shown by the large number of transitions between different categories.

Our visualization tool was developed using the D3 JavaScript library. The aim of this visualization is to investigate temporal trajectories of HbA1c levels for a large cohort of diabetes patients over a number of years prior to death. Parallel sets is chosen for showing HbA1c summary trajectories. Each vertical axis is a time step. The user can choose the frequency of these time steps, with a minimum sampling frequency of six months, and also the maximum number of years before death. The death event axis is placed at the right with all other time steps moving backwards in time to the left (Figure 1). Each vertical axis is split into the four HbA1c categories (Normal in green, Borderline in blue, Controlled in orange, and Uncontrolled in red), and a Missing category in grey (for patients with more than 10 years of data). The height of each axis category represents the proportion of the patients in that category at that point in time. Paths moving between axes recursively split, moving backwards from death to show the trajectories of similar groups of patients. The visualization can show trends either starting at the death event, i.e. going backwards

in time, or starting at the last year in the visualization, i.e. going forward in time. The user can highlight one or more groups of patients by clicking on categories or trajectories to highlight the behavior of that group of patients going backward and forward in time, reducing visual clutter (Figure 2). We also include the ability to incorporate demographic data, such as gender, as additional axes (Figure 1). This feature enables the comparison of trajectories for different subpopulations based on data other than just HbA1c levels.

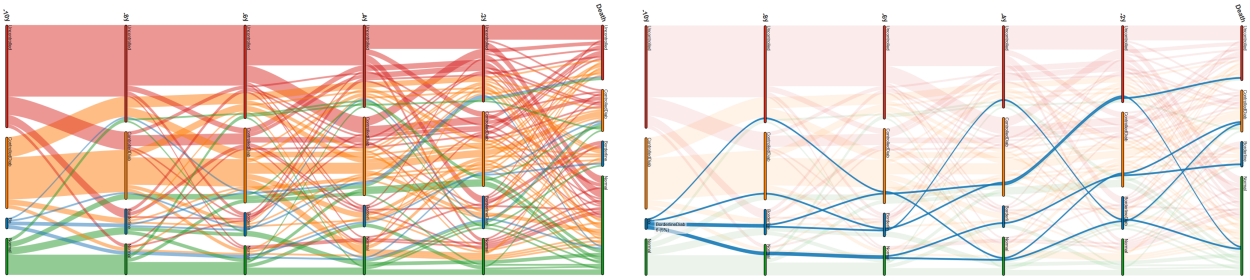


Figure 2. A 10-year range of data, sampled every two years, with forward propagation to show how the trajectories of patients change moving forward in time (left). Highlighting enables a focused view of a single category, reducing visual clutter (right).

The user can also chose between different types of coloring schemes for the paths: 1) color by the category at the first or last year (depending on the propagation direction), which shows the level of variation for a category over the length of the visualization; 2) color by transition, where the transition has a gradient from the source to target category color, useful for showing overall trends; and 3) color by reverse transition, where the transition path has a gradient from the target category to the source category, useful for category-level analysis of the distribution of source and target categories at a particular time step's category (Figure 3). To reduce visual clutter, there is also an option to look at only static transitions (i.e. no change in category between time steps), and to look at only variations (i.e. only changes in the categories).

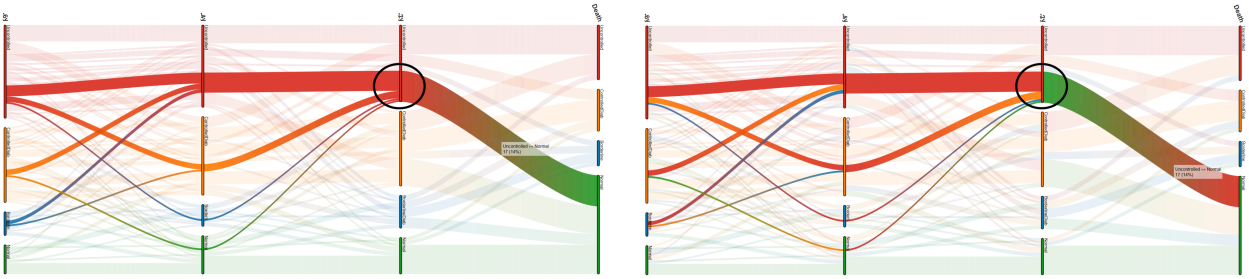


Figure 3. In addition to coloring by the starting category, paths can be colored by a gradient from source to target category (left), which redundantly encodes the category at each axis to emphasize overall trends, or by target to source category (right), which enables a rapid analysis of where paths are moving to/from at each category. The circled regions highlight this difference. On the right, it is immediately obvious what category this trajectory came from at death (Normal in green) and how this group is distributed at the previous time step.

Acknowledgments

This work is supported by the US Army Medical Research and Materiel Command (USAMRMC) under Grant No. W81XWH-13-1-0061. The views, opinions and/or findings 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. We acknowledge the assistance from Meghana Ganapathiraju in helping to refine the visualization. Our implementation was adapted from the d3.parsets reusable chart by Jason Davies.

References:

1. Gebregziabher M, Egede LE, Lynch CP, Echols C, Zhao Y. Effect of trajectories of glycemic control on mortality in type 2 diabetes: a semiparametric joint modeling approach. Am J Epidemiol. 2010;171(10):1090-1098
2. Bendix F, Kosara R, Hauser H. Parallel sets: visual analysis of categorical data. IEEE Symp on Info Vis. 2006;12(4):133-140.
3. Wongsuphasawat K, Gotz D. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. IEEE TransVis Comput Graph, 2012;18(12):2659-2668.

Interactive Event Sequence Visualization and Querying

EventFlow Demonstration

Fan Du, Sana Malik, Catherine Plaisant, Ben Shneiderman
University of Maryland
Human-Computer Interaction lab and Department of Computer Science
Contact: {fan, maliks, plaisant, ben}@cs.umd.edu
www.cs.umd.edu/hcil/eventflow

The University of Maryland Human-Computer Interaction Lab has developed EventFlow, which has distinctive innovative features such as: (1) visual representation of temporal patterns (of point and interval events) for individual records and for the aggregate of all records (2) novel graphical query language to pose temporal queries such as Find all records with Event A followed by B followed by C (infection, fever, bleeding), or Event A during Interval B (stroke while taking warfarin), (3) query results are presented visually, organized by the matching and non-matching histories, (4) a search and replace feature (replace all sequences of normal blood pressure point events with an interval that shows the duration, (5) efficient internal data structures to support retrieval and aggregated view presentation, and (6) many features to organize, clean, transform, and simplify the data.

Temporal event data is a fundamental component of electronic health records. As such, many visualization tools have been designed for the exploration of this data type, however, they rely heavily on the assumption that the underlying data is fit to be explored. In many cases though, event patterns must be extensively transformed in order to better reflect either the real world events that generated them or the perspective of a given study. Without this step, population-level trends can be obscured.

Temporal event data wrangling, however, is deceivingly difficult and error prone even for expert users. Standard, command-based query languages are poorly suited for specifying even the simplest event patterns, and attempts at more accessible query languages frequently omit critical features such as events that occur over a period of time (intervals) or the absence of an event. Perhaps most importantly is that query alone is not enough to get users through a typical temporal event data wrangling process. Event patterns not only need to be found, but also transformed and re-represented. An improved query and wrangling process not only benefits database professionals, but also dramatically increase the range of users who can access this type of data.

The EventFlow visualization tool is built to extend beyond the typical bounds of data exploration, and serve as a critical aid for both temporal event query and data transformation.

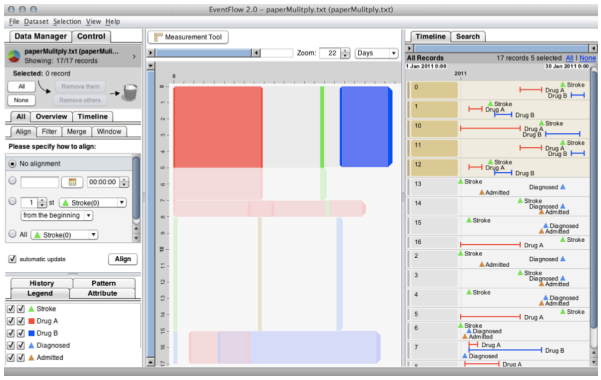


Figure 1 – The EventFlow interactive analysis tool (www.cs.umd.edu/hcil/eventflow) with a small sample dataset. On the left are found controls and legend, in the middle is the overview of all sequence patterns in the dataset, and on the right a scrollable timeline browser shows all the individual records. The top sequence in the overview is selected (drug A, followed by stroke, followed by drug B). The distance between events corresponds to the average time between events. The height of the bar corresponds to the proportion of records with that sequence. The records with the selected sequence are highlighted at the top in the timeline view.

ACKNOWLEDGEMENTS

EventFlow is supported by Oracle Corporation, The Center for Health-Related Informatics and Bioimaging (CHIB) at the University of Maryland, the Maryland Industrial Partnerships (MIPS) program and Pulse8.

PAPERS

1. Monroe, M, Lan, R., Lee, H., Plaisant, C., Shneiderman. B., Temporal Event Sequence Simplification. *IEEE Conference on Visual Analytics Science and Technology (VAST* 2013), 2227-2236.
2. Monroe, M., Lan, R., Morales del Olmo, J., Shneiderman, B., Plaisant, C., Millstein, J., The Challenges of Specifying Intervals and Absences in Temporal Queries: A Graphical Language Approach. *Proc. of ACM Conference on Human-Computer Interaction* (CHI 2013), 2349-2358.

HCC Risk Browser: Visualizing Opportunities and Interventions

Michael A. Simon, PhD[1], Nick C. W. Stepro[1]
[1]Arcadia Healthcare Solutions, Burlington, MA

Abstract

Effective extraction of information related to patient care has only grown in importance with increased demands on a more productive provider environment, more effective population health management, and lower overall cost of care for a diverse population. Yet, visibility into even basic measures is inhibited by disparate and inadequate storage systems, communication mechanisms, and provider workflows[1]. One such example is the Centers for Medicare and Medicaid Services (CMS) Hierarchical Condition Category (HCC) Risk Adjustment algorithm. This algorithm is canonically defined by one data stream but is in fact better defined by a much wider dataset. Here, we describe a data model meant to visualize sources and documentation of risk, designed with a specific group/zone visualization intent, rather than to reporting or specific display requirements. This model, layered on a source-agnostic data warehouse with proficient master patient index, facilitated the identification and documentation of thousands of documentation and care gaps among a sample population, at little marginal cost to providers. We will also show the power of interactive and animated data exploration, and will pose a challenge for visualization experts in this climate to advance the entire healthcare ecosystem to a new standard for unearthing and transferring insights, one in which data models are designed not by what is at hand but rather by what can be learned.

Visualization

The HCC Risk Browser (Figure 1) came out of a need by healthcare payers and provider networks to better understand member-patient risk in the context of the CMS HCC Risk Adjustment Algorithm.

While CMS makes use of information reported by way of the Risk Adjustment Processing System (RAPS), highly reduced claims-style records submitted by the payer, anyone wishing to understand or validate these results must seek out a number of sources, including:

(1) Claims from all patient encounters

(2) EHR information, including encounters, vitals, progress notes, medications, and lab orders

(3) CMS RAPS, model output, and member reports.

By integrating these data sources, and searching out both retrospective opportunities for risk adjustments as well as prospective opportunities for proactive interventions, the HCC Risk Browser visualizes patient chronic conditions in an interactive and actionable format. Aggregate results in total and by HCC code or category are presented clearly, and with easy interactivity by the user (Figure 1, top).

To better visualize opportunities on a patient-by-patient level, patients are represented in a modified pareto-figure as discs in order of total identified risk of chronic condition (Figure 1, bottom; and Figure 2). The discs are colored by the opportunity group represented in order to the user to specifically focus on any of (a) the patient with the

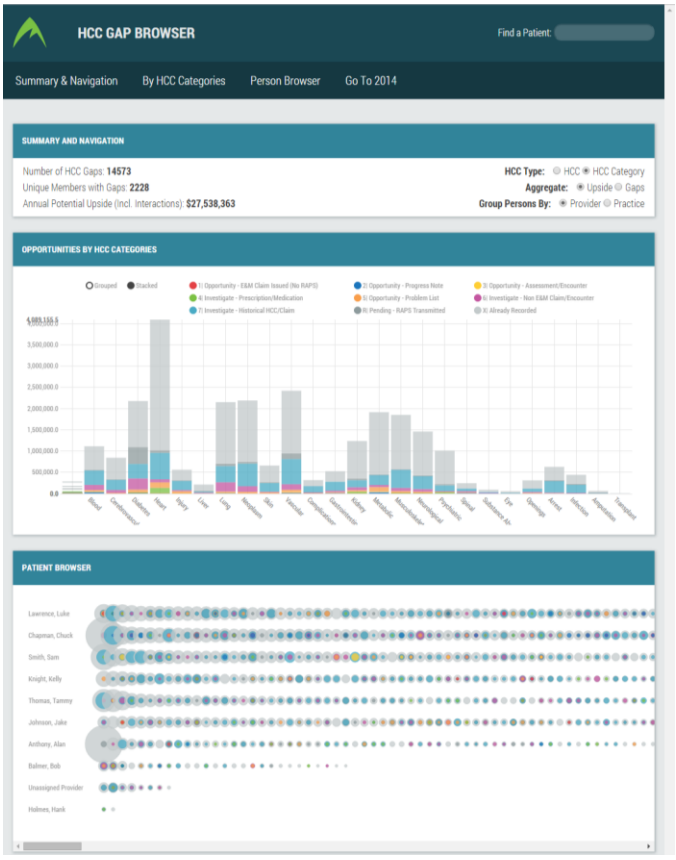


Figure 1. Interactive HCC Gap Browser presents information organized by HCC category, opportunity groupings, and patient-level opportunity risk and risk gaps.

Table 1. Organization of Risk Identification and Intervention Opportunities

X — ***Already Recorded in MOR***
CMS' Model Output Report already reflects the HCC group

R — ***Pending RAPS Transmission***
CMS RAPS Return File reflects encounters in the HCC group

1 — ***Claim Issued, no MOR/RAPs***
A claim exists for the encounter, but no record exists in RAPS Return Files
This group may be recovered by identifying and (re-) submitting the matching claim.

2 — ***Signed Progress Note***
A signed progress note contains an assessment in the HCC group
Documentation of the encounter makes recovery of this group straightforward

3 — ***Face-to-Face Visit w/ Assessment***
A record of an encounter exists in the EHR with a relevant diagnoses, but not incorporated to a signed progress note.
Provider can close visit and submit claim.

4 — ***Medications or Prescriptions***
A pharmacy claim or medication linked to an encounter indicates the presence of a condition
This group may indicate a care gap or documentation disconnect

5 — ***Active Problems on Problem List***
An entry in the Problem List may indicate the presence of a ongoing or chronic condition
Presence in this group may reveal care gaps and/or documentation disconnect

6 — ***Claim Issued w/o E&M Code***
Claims issued in the current year for which there is no face-to-face office-visit encounter
This group may reveal care gaps and ongoing poor maintenance of chronic conditions

7 — ***Historical Claim Issued***
A condition was signaled by a claim or CMS report in a previous year
The historical presence of conditions may reveal care gaps in the measurement year and opportunities for improved care in the present

greatest chronic condition risk, (b) a specific opportunity or intervention group represented within the disc, or (c) the patients assigned to a specific PCP.

By selecting a single individual, the user is presented with a list of opportunities and interventions identified within that individual (Figure 3). In the hands of a plan risk manager, this information can dramatically improve efficiency of risk reporting and capture efforts. In the hands of a clinician, this information offers a shorthand for questions about current or historical health conditions and wellness practices. In the hands of a care coordinator, this information can facilitate patient outreach for proactive condition management and appropriate preventative care.

Finally, no solution is complete without transparency and verifiability. By selecting a specific HCC category, the user has access to an "evidence table", which lists the various references within all available data sources (claims, clinical, and RAPS) to substantiate the risk determination (Figure 4). By stepping through this information, the user can act immediately on references, actions that may include requesting immediate recognition and reimbursement from CMS, seeking out potentially outdated information for update, identifying providers or specialists for additional documentation, and contacting patients overdue for proper treatment of a chronic condition.

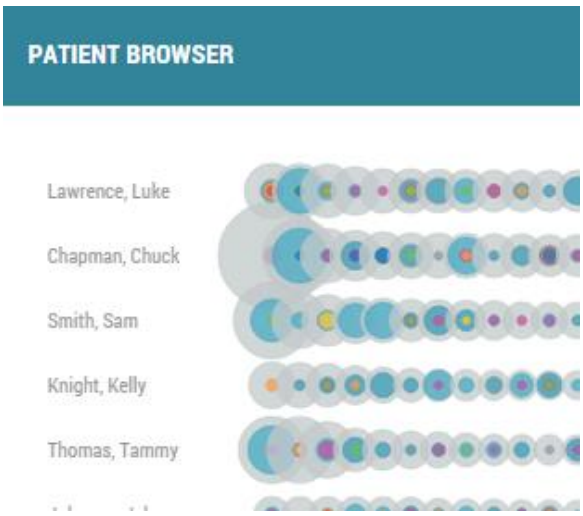


Figure 2. Patient Browser view presents patients in order of risk identified, color-coded to opportunity and intervention group (See Figure 1 for color key).

Results

An initial small-scale pilot was run on a large payer organization in the Northwest United States, and a full pilot was then run on over 3,300 member-patients at a managed care organization in the Southeast United States. Based on initial results from these trials, retrospective analysis identified hundreds of unreported risk conditions, stemming

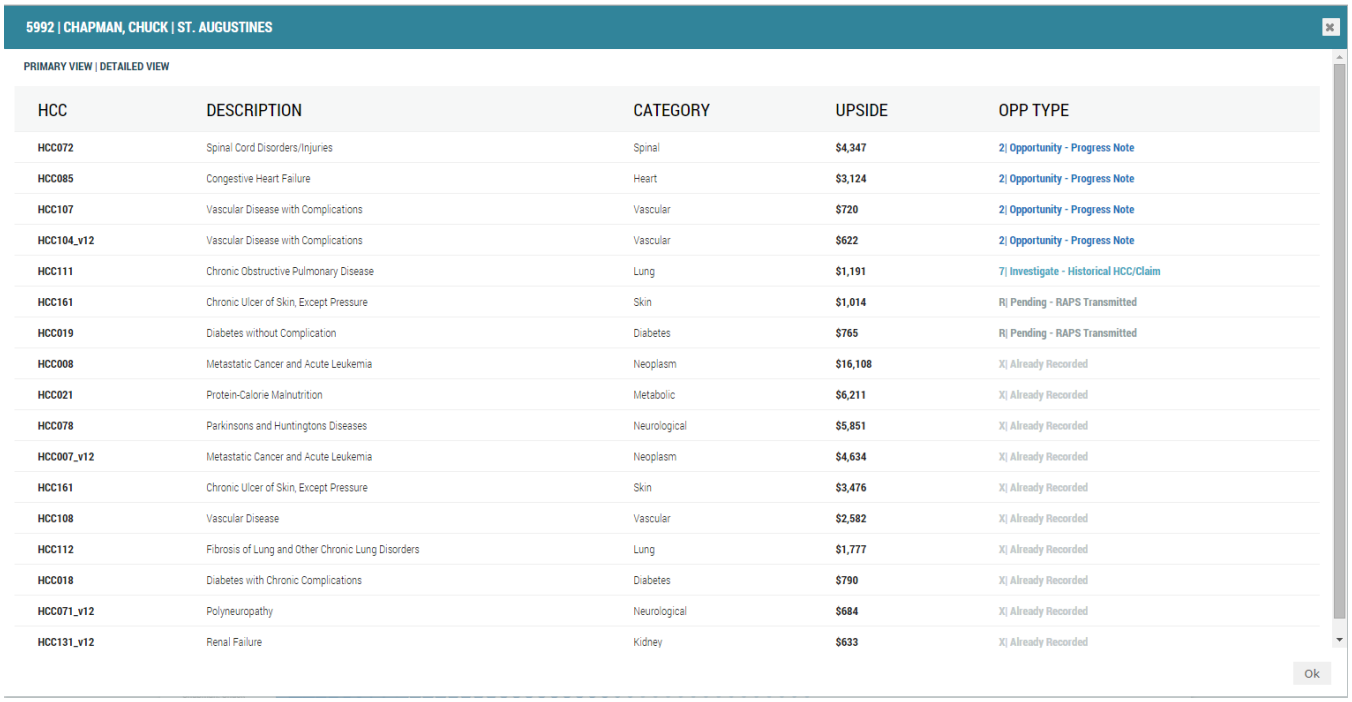


Figure 3. Overview of risk opportunities and interventions identified for a specific patient, by category and group.

from to claims reporting errors, specialist documentation gaps, and failure to link assessments within EHRs. The resulting risk adjustments by CMS potentially represent between $100 and $200 per member per year in retroactive reimbursement.

Furthermore, prospective analysis of those 3,300 member-patients revealed thousands of potential intervention opportunities to cover documentation or care gaps. These activities not only create a more accurate representation of the overall risk of an organization's member-patient population. They also improve healthcare for the entire community by directly identifying potential care gaps for intervention, as well as identifying processes that could contribute to care and documentation gaps in the future.

Conclusion

Peering into the CMS HCC Risk Adjustment algorithm represents a challenge of multi-dimensional, heterogeneous datasets with mixed availability and visibility. By addressing this challenge with a data model designed around effective visualization, and with the aid of intuitive multi-dimensional, interactive data navigation tools, we have developed a technique that provides depth, breadth, verifiability, and functionality of insight into a complex clinical and healthcare challenge.

References

1. Simon MA, Baum Z, Lebel L, Harvey L, Gillis B. Mind the Gap: Identifying critical data quality gaps to unlock population health management. J. Healthcare Inf. Management. 2014; 28(2): 28-33.

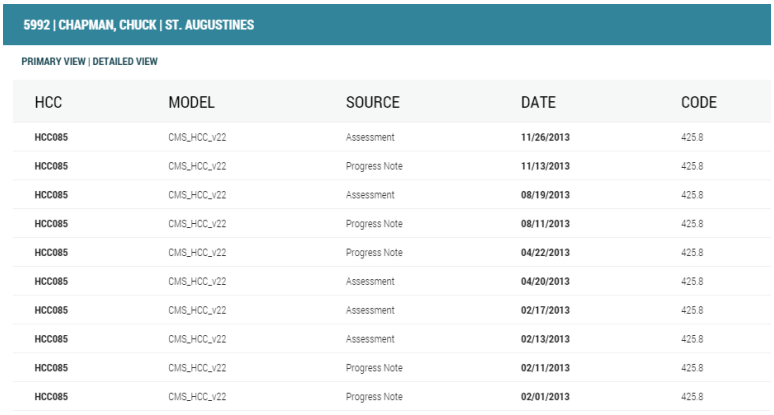


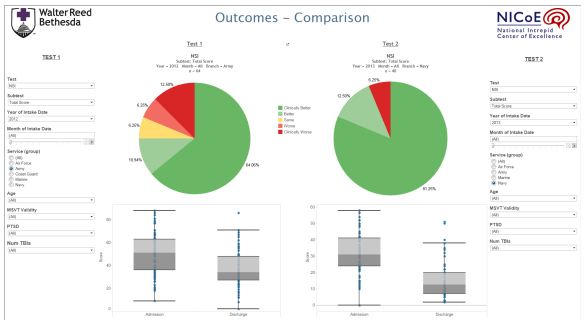
Figure 4. Evidence table, presenting all references related to a specific risk opportunity or intervention.

Understanding Outcome Measures of Patients Diagnosed with mild TBI - A Visual Analytics Approach

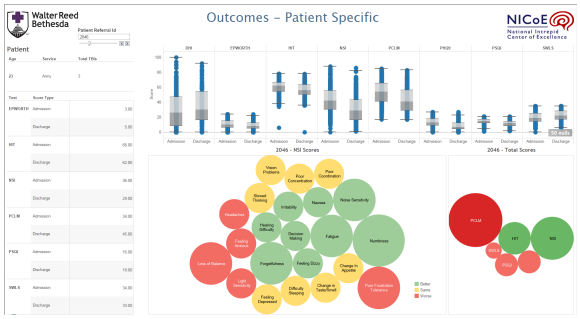
Ryan Diehl, Niki Noprapa, and Jesus J Caban

National Intrepid Center of Excellence (NICoE),
Walter Reed National Military Medical Center, Bethesda, MD

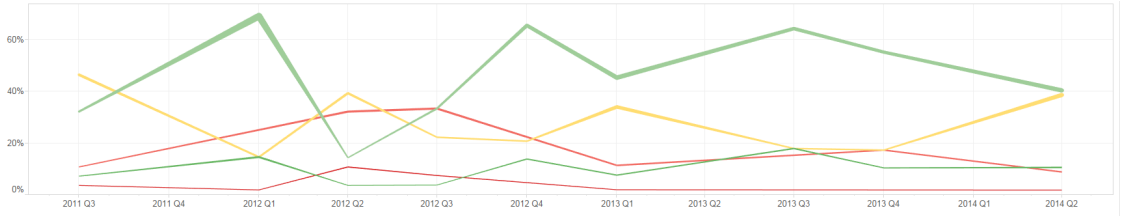
Assessing, analyzing, and understanding short- and the long-term clinical outcomes of patients that have been diagnosed with mild Traumatic Brain Injury (mTBI) is a challenging tasks. There are a number of potential metrics, surveys, and clinical elements that could be used to measure progression and response to treatment. Unfortunately the large number of potential variables that can be used to assess progression makes it very challenging for providers, clinical staff, and researchers to understand the results of brain injuries and the effects of different treatment modalities. We present a flexible and easy-to-use visual analytics software application that can be used by either researchers or clinicians to quickly analyze outcomes data. The clinical dashboard provides the user with the ability to compare and contrast multiple outcome surveys at the same time. Users can select any combination patient tests/subtests, year/month, branch, age group, validity, PTSD diagnosis, and number of TBIs and compare the results of different cohorts side-by-side. The system allows for tending of the data as well as the analysis of each patient individually.



Side-by-side comparison of different cohorts.



Patient specific outcomes data can be display to better understand individual symptoms



Trending of the data is supported to better analyze the overall response to treatment.

Visual Insight for Better Decision: Revealing Meaningful Values of Visual Analytics in Healthcare Dashboards

Yair G. Rajwan, DSc, MS, Visual Science Informatics

Accurate review of a patient's chief complaint, history of present illness, and medical history is a critical antecedent to a medical diagnosis and the development of a treatment plan. Electronic medical systems can capture and display a dashboard of a patient's medical history items on a chronological timeline. However, temporal visualization, in healthcare dashboards, mostly focuses on displaying objective and quantitative information that a healthcare practitioner measures from a patient or coded results from laboratory tests and diagnostics. While a patient's chief compliant and history of present illness are subjective components and captured in a narrative form.

Given the free-text form of clinical notes, text visualization can be explored in displaying subjective and qualitative information that a patient communicates to a healthcare practitioner such as chief complaint, history of present illness, medical history, family history, as well as social history. At the same time, the text visualization must be combined with temporal visualization to depict a progression of clinical notes sequence.

This demonstration illustrates a visual text analytics system to support a hybrid of text visualization and temporal visualization and applies this prototype to a single patient's clinical notes.

To be used at the point-of-care, the hybrid visualization shows a single patient's current incidents and compares them to past events. The events comparison includes views of new, vanished, clustered and singular sign, symptom, diagnostic, test, or treatment.

Biographical Sketch

Yair G. Rajwan, DSc

Dr. Yair Rajwan is the founder and director of Visual Science Informatics, a Virginia firm that helps organizations analyze and visualize data and text to provide insight and improve engagement. He applies open source platforms to harmonize organizations' datasets with open data, analyze data and text, design information visualization, and develop visual analytics interfaces. He is the founder of VisualMatics - Visual Science Informatics, a social network group that connects healthcare practitioners, visual analytics professionals, and information visualization researchers to collaborate on visual analytics proposals and grants. His enthusiasm to make a different in healthcare systems started in 2009 by designing and evaluating patient-oriented visualization to improve healthcare outcomes. His research was published as part of his postdoctoral fellowship of the National Library of Medicine at the Division of Health Sciences Informatics, Johns Hopkins University School of Medicine. His published contribution was visualizing infections outcome data to health care consumers and practitioners for decision making. Its impact, at the Maryland Center for Hospital Services, was reduced infections in eight hospitals. His peer-review publications, conference presentations, and educational materials are listed at the Visual Science Informatics Forum.

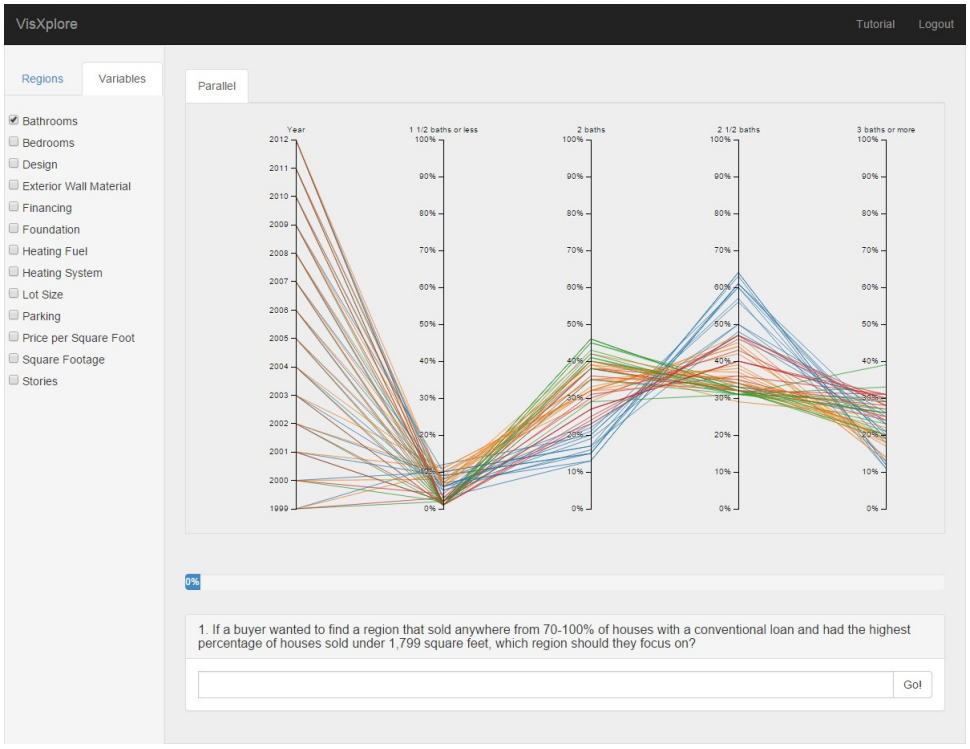
An Interactive Visualization System with a Grammar Induction Layer for Learning and Generating Suggestions from Complex Clinical Datasets

Filip Dabek[1], Jesus J Caban[2] and Tim Oates[1]

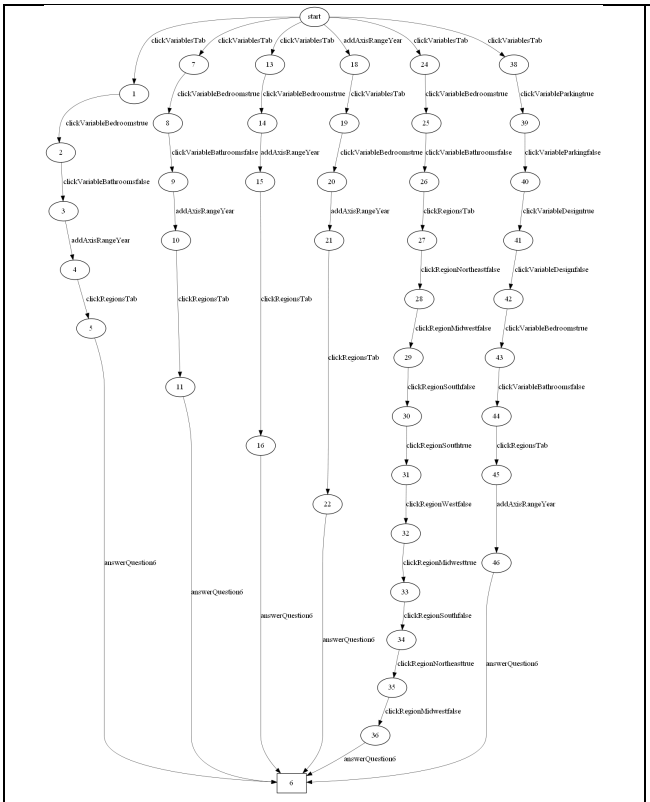
[1]University of Maryland, UMBC

[2]National Intrepid Center of Excellence (NICoE), Walter Reed Bethesda

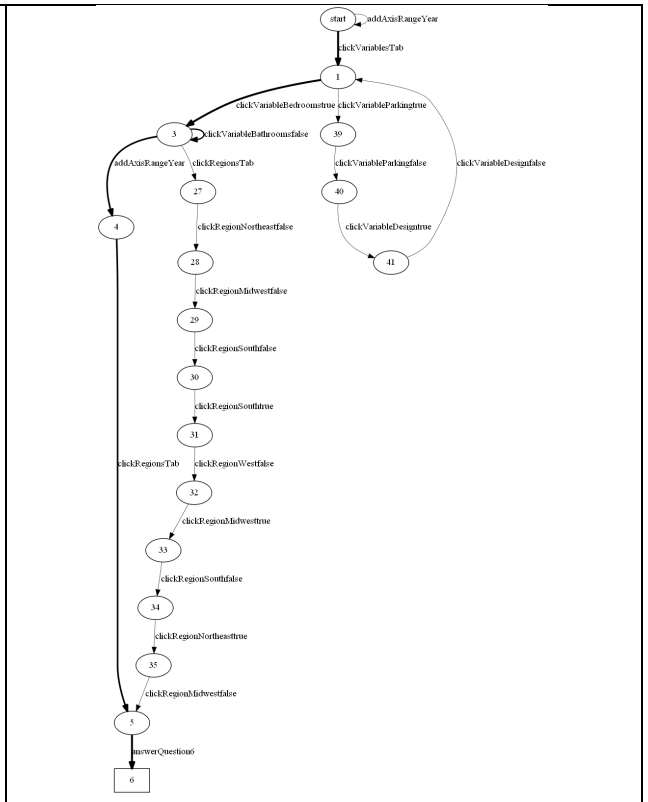
Assessing, exploring, and analyzing large amount of heterogeneous clinical data are some of the key challenges faced by researchers and clinicians interested in identifying complex clinical patterns. Flexible visualization systems capable of performing individual statistical analysis while guiding the user towards patterns in a data set have the potential of impacting the way clinical datasets are analyzed. We introduce VisXplore -- a clinical data visualization system that has been designed to perform a wide range of analysis techniques on clinical data, along with a machine learning component aimed at assisting users in obtaining answers to their clinical questions through the use of interaction suggestions. The visualization system captures and aggregates interaction data from users and uses grammar induction along with a set of pre-defined rules to identify a lost user and keep them on track in obtaining an answer from a visualization. The system has been tested with different clinical datasets, while our suggestions have been tested and shown to produce equal or significantly better performance in users 91.67% of the time, concluding that we are able to guide a user along the visual analytic process.



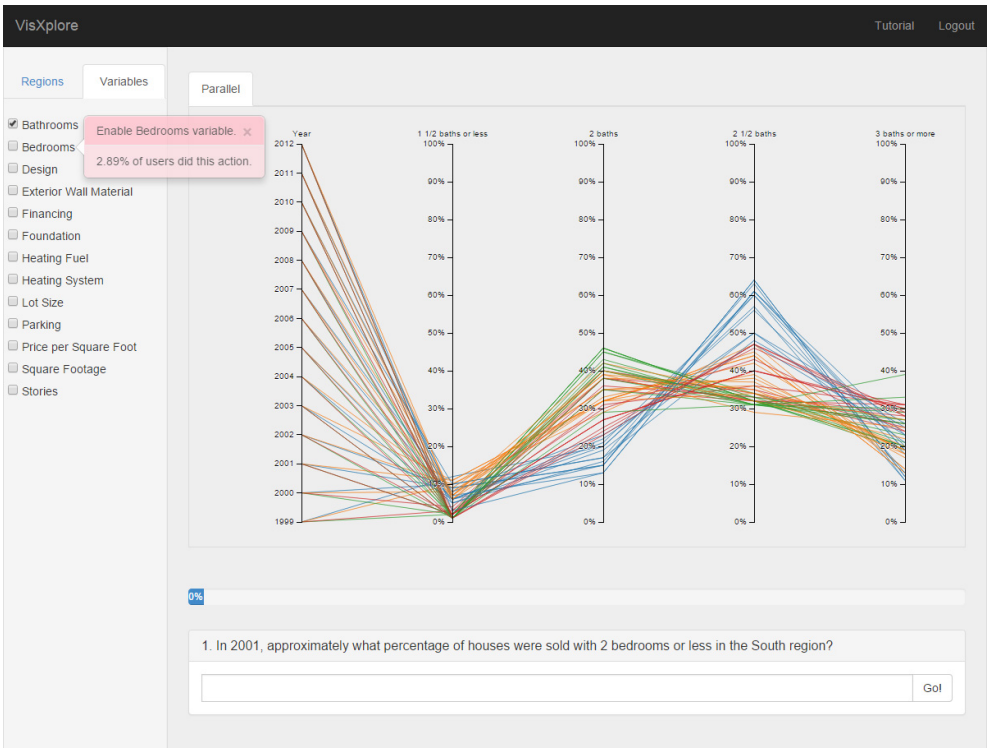
User Interface for our Visualization System, VisXplore



User Interaction Graph of 6 Users



Compressed User Interaction Graph



User Being Presented with a Suggestion Generated from the Compressed Graph

Program Committee

Organizers:

Jesus J Caban, PhD
National Intrepid Center of Excellence (NICoE),
Walter Reed Bethesda
Email: jesus.j.caban.civ at mail.mil

David Gotz, PhD
University of North Carolina
Email: gotz at unc.edu

Adam Perer, PhD
IBM T. J. Watson Research Center
Email: adam.perer at us.ibm.com

Hadi Kharrazi, MD, PhD
Johns Hopkins Bloomberg School of Public Health

Program Committee	
Michael J. Ackerman, PhD	National Library of Medicine
Patti Brennan, PhD, RN	University of Wisconsin-Madison
Ketan Mane, PhD	University of North Carolina, Chapel Hill
Yair Rajwan, PhD	Visual Science Informatics
Paul G. Nagy, PhD	Johns Hopkins University
Catherine Plaisant, PhD	University of Maryland, College Park
Suzanne Bakken, PhD, RN	Columbia University
Terry S. Yoo	National Library of Medicine