VAHC 2011

Proceedings of the IEEE VisWeek

Workshop on Visual Analytics in Healthcare:

Understanding the Physicians Perspective

October 23rd, 2011
Providence, RI
www.visualanalyticshealthcare.org

Sponsors:

NATIONAL INSTITUTES OF HEALTH

NATIONAL LIBRARY OF MEDICINE

Preface

Visualization and visual analytics show great potential as methods to analyze, filter, and illustrate many of the diverse data used in clinical practice. Today, (a) physicians and clinical practitioners are faced with the challenging task of analyzing large amount of unstructured, multi-modal, and longitudinal data to effectively diagnose and monitor the progression of a particular disease; (b) patients are confronted with the difficult task of understanding the correlations between many clinical values relevant to their health; and (c) healthcare organizations are faced with the problem of improving the overall operational efficiency and performance of the institution while maintaining the quality of patient care and safety.

Visualization and visual analytics can potentially provide great benefits to each of these three core areas of healthcare. However, to be successful, the resulting visualization must be able to meet the physician's requirements and be useful for both patients and physicians.

Despite the continuous use of scientific visualization and visual analytics in medical applications, the lack of communication between engineers and physicians has meant that only basic visualization and analytics techniques are currently employed in clinical practice. The goal of this workshop is to gather together leading physicians and clinical practitioners to share with the visualization community their need for specific visualization tools and discuss the areas in healthcare where additional visualization techniques are needed.

Jesus J Caban,
NICoE / Naval Medical Center
CC / National Institutes of Health

David Gotz
IBM Research

Invited Speakers

Dr. Joe Terdiman, MD, PhD
Kaiser Permanente Division of Research

Joe Terdiman, MD, PhD, is a research scientist at the Kaiser Permanente Northern California Division of Research; and an assistant professor in the School of Optometry, University of California, Berkeley. He has been with the Division of Research since 1969, first as a medical information scientist, and then as an assistant to the director prior to becoming a research scientist. Dr. Terdiman's research interests include medical informatics, technology assessment, cardiovascular epidemiology, and cardiovascular and visual physiology. Dr. Terdiman is the principal investigator of The Kaiser Permanente National Research Database.

Dr. Jeffrey L. Schnipper, MD, M.P.H
Harvard Medical School / Partners HealthCare System
Dr. Schnipper is an Assistant Professor of Medicine at Harvard Medical School, Associate Physician at Brigham and Women's Hospital (BWH), and Director of Clinical Research for the BWH Hospitalist Service.

His research interests focus on improving the quality of health care delivery for general medical patients. Subject areas include preventive cardiology, inpatient diabetes care, safe and effective medication use, transitions in care, and communication among health care providers. The quality improvement interventions that he studies include the greater use of information systems, hospital-based pharmacists, and process redesign using continuous quality improvement methods.

Paul Nagy, PhD
Johns Hopkins University
Dr. Nagy is a Visiting Associate Professor and Director of Quality at the Russell H. Morgan department of Radiology at Johns Hopkins University. Dr. Nagy research interests include utilizing information technology as a platform to measure quality in radiology.

In 2010 he became the chair of the American Board of Imaging Informatics (ABII). ABII is a society formed in 2007 by the American Registry of Radiologic Technologists and the Society of Imaging Informatics in Medicine

Agenda:

Session I: Frameworks for Visual Analytics in Healthcare Chair: David Gotz, IBM Research	
8:30 - 8:40	Welcome
8:40 - 9:20	Paper, Posters, and Demos Fast Forward -- Flash Presentations
9:20 - 10:10	Paper presentations
	"VisCareTrails: Visualizing Trails in the Electronic Health Record with Timed Word Trees, a Pancreas Cancer Use Case" Lauro Lins, Marta Heilbrun, Juliana Freire and Claudio Silva
	"AnamneVis: A Framework for the Visualization of Patient History and Medical Diagnostics Chains" Zhiyuan Zhang, Faisal Ahmed, Arunesh Mittal, Iv Ramakrishnan, Rong Zhao, Asa Viccellio and Klaus Mueller
	"Engaging Clinicians in the Visualization Design Process - Is It Possible?" Kostas Pantazos
10:10 - 10:30	Coffee Break
Session II: Physicians @ VisWeek Chair: Jesus J. Caban, NICoE / Naval Medical Center & NIH	
10:30 - 11:10	**Keynote:** Dr. Jeffrey L. Schnipper, MD, M.P.H Brigham and Women's Hospital / Partners Healthcare
11:10 - 12:10	**Panel**: Physicians @ VisWeek
	Dr. Joe Terdiman, MD, PhD Kaiser Permanente
	Dr. Jeffrey L. Schnipper, MD, M.P.H Brigham and Women's Hospital
	Dr. Paul Nagy, PhD Johns Hopkins University
	Dr. Marta Heilbrun, MD, MS University of Utah
12:10 – 1:30	Lunch break

Session III: Visualizing EMR data Chair: Klaus Mueller, Stony Brook University	
1:30 - 2:10	**Keynote:** Dr. Joe Terdiman, MD, PhD Kaiser Permanente Division of Research
2:10 – 3:00	Paper presentations
	"Outflow: Visualizing Patients Flow by Symptoms and Outcome" Krist Wongsuphasawat and David Gotz
	"Clinical Applications of Start Glyphs and Ideas about Crowdsourcing Data Visualization Software" Jim DeLeo and James J Cimino
	"*Visual Interactive Quality Assurance of Personalized Medicine Data and Treatment Subtype Assignment"* Edward Worbis, Raghu Machiraju, Christopher Bartlett and W. Ray
3:00 – 3:45	Poster and Demo Session
Session IV: Visual Analytics - Beyond EHR Visualization Chair: James DeLeo, National Institutes of Health (NIH)	
3:45 – 4:30	**Keynote**: Dr. Paul Nagy, PhD Johns Hopkins University
4:30 – 5:15	Paper presentations
	"*Hierarchical Summarization of Concepts for Visual Discovery Browsing - a Pilot Study*" Michael J Cairelli and Thomas C. Rindflesch
	"Assessing Risks for Families with Inherited Cancers" Brian Drohan, Curran Kelleher, Georges Grinstein and Kevin Hughes
	"Interactive Visualization for Understanding and Analysing Medical Data" Samar Al-Hajj, Richard Arias and Brian Fisher
5:15	Closing Remarks

Poster Presentations (3:00 – 3:45pm)

"Trauma Analysis through Data-Driven Medical Injury Visualization"
Patrick Gillich

"Quantitating pathogenic biofilm architecture in biopsied tissue"
Shareef Dabdoub, Brian Vanderbrink, Sheryl Justice and William Ray

"TAO: Terrain Analytic Operators for Expert-Guided Data Mining Applications",
Jason Mclaughlin, Qian You, Shiaofen Fang and Jake Y. Chen

"The (Inter)face of Kalm",
Halimat Alabi, David Worling and Bruce Gooch

Demo Presentations (3:00 – 3:45pm)

"BodyTrack: Open Source Tools for Health Empowerment through Self-Tracking"
Anne Wright and Ray Yun

"VisCareTrails: Visualizing Trails in the Electronic Health Record with Timed Word Trees, a Pancreas Cancer Use Case"
Lauro Lins, Marta Heilbrun, Juliana Freire, and Claudio Silva

"Food For The Heart: Visualizing Nutritional Contents for Food Items for Patients with Coronary Heart Disease"
Fransisca Vina Zerlina, Bum chul Kwon, Sung-Hee Kim, Karen S. Yehle, Kimberly S. Plake, Sibylle Kranz, Lane M. Yahiro, and Ji Soo Yi

"ImageVis3D Mobile in Clinical Use"
Jens Kruger and Thomas Fogal

"AnamneVis: A Framework for the Visualization of Patient History and Medical Diagnostics Chains"
Zhiyuan Zhang, Faisal Ahmed, Arunesh Mittal, IV Ramakrishnan, Rong Zhao, Asa Viccellio, and Klaus Mueller

"LifeFlow: Understanding Millions of Event Sequences in a Million Pixels"
Krist Wongsuphasawat

"Hierarchical Summarization of Concepts for Visual Discovery Browsing"
Michael J. Cairelli and Thomas C. Rindflesch

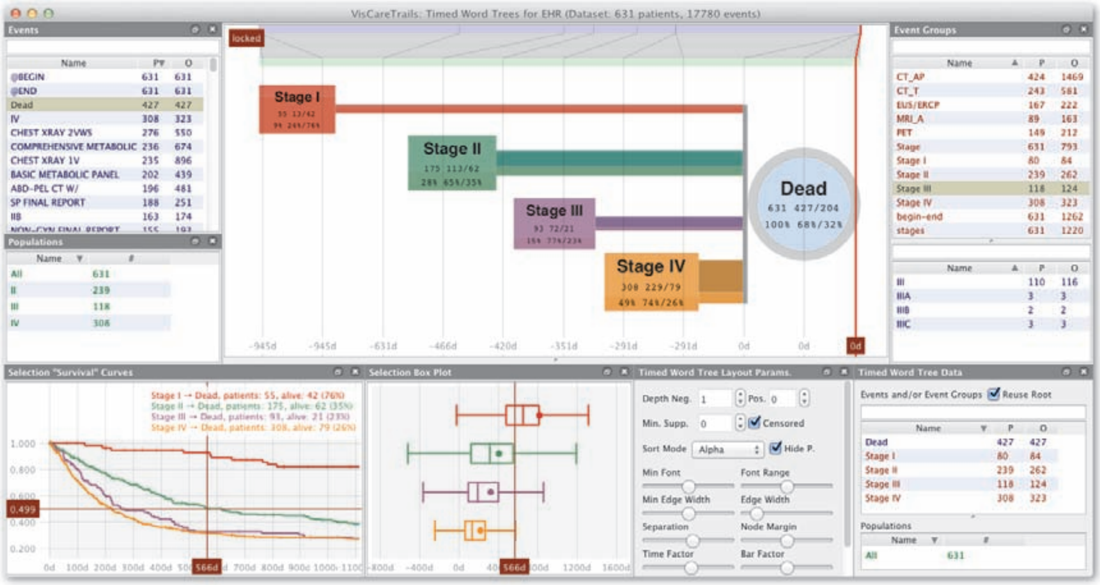
"InBox: In-situ Multiple-Selection and Multiple-View Exploration of Diffusion Tensor MRI Visualization"
Jian Chen, Haipeng Cai, Alexander P. Auchus

Proceedings of the IEEE VisWeek Workshop on Visual Analytics in Healthcare:

Understanding the Physicians Perspective

VisCareTrails: Visualizing Trails in the Electronic Health Record with Timed Word Trees, a Pancreas Cancer Use Case

Lauro Lins, Marta Heilbrun, Juliana Freire, *Member, IEEE*, and Claudio Silva, *Member, IEEE*



Events

Name	P	O
@BEGIN	631	631
@END	631	631
Dead	427	427
IV	308	323
CHEST XRAY 2VWS	276	550
COMPREHENSIVE METABOLIC	236	674
CHEST XRAY 1V	235	896
BASIC METABOLIC PANEL	202	439
ABD-PEL CT W/	196	481
SP FINAL REPORT	188	251
IIB	163	174
NON-GYN FINAL REPORT	155	193

Populations

Name	#
All	631
II	239
III	118
IV	308

Event Groups

Name	P	O
CT_AP	424	1469
CT_T	243	581
EUS/ERCP	167	222
MRI_A	89	163
PET	149	212
Stage	631	793
Stage I	80	84
Stage II	239	262
Stage III	118	124
Stage IV	308	323
begin-end	631	1262
stages	631	1220

Name	P	O
III	110	116
IIIA	3	3
IIIB	2	2
IIIC	3	3

Timed Word Tree Layout Params.

Depth Neg. 1 Pos. 0
Min. Supp. 0 ☑ Censored
Sort Mode Alpha ☑ Hide P.

Timed Word Tree Data

Events and/or Event Groups ☑ Reuse Root

Name	P	O
Dead	427	427
Stage I	80	84
Stage II	239	262
Stage III	118	124
Stage IV	308	323

Populations

Name	#
All	631

Fig. 1. VisCareTrails session on a dataset of pancreatic cancer patients. The central top display shows a *Timed Word Tree* with staging events (Stage I, Stage II, Stage III, Stage IV) and rooted in the death event (Dead). Selecting the stage nodes, corresponding to severity and extent of disease, the bottom left plot presents survival curves indicating the fraction of each of the four sets of staged subjects that were still alive after t days, and the box-plot represents the distribution of the time distance the death event. This visualization confirms that this specific dataset follows the known patterns for pancreatic cancer patients and is obtained with just a few intuitive mouse gestures.

Abstract— As a mandate in the 2009 ARRS act, all US health care systems are moving toward electronic health record (EHR) systems to capture and store patient data. The EHR is a rich source of health information about individual patients and/or populations. The ability to analyze and identify meaningful patterns in this data has the potential to produce important knowledge. Yet, there is still a considerable gap between what answers are captured in this record and what answers can be effectively extracted from it. To reduce this gap, more intuitive ways of posing questions and obtaining answers are needed. In this paper we present VisCareTrails, a system based on *timed word trees* visualization that summarizes event paths relative to a given *root event* and are obtained through a simple drag-and-drop user interface. These summaries visually convey information about the nature, frequency and average timing of the event paths, and serve as a natural starting point to obtain further details and compare different paths. We apply VisCareTrails in a dataset of pancreatic cancer patients to illustrate its effectiveness.

Index Terms— Information visualization, Electronic Health Records, Survival, Cancer, Word Trees, Tree Layout.

1 Introduction

As a component of the ARRS and HITECH acts of 2009, the US government has made a significant investment in order to grow the Electronic Health Record (EHR). Hospitals and providers who demonstrate “meaningful use” of the EHR will begin receiving incentive payments in 2011, with penalties to begin after 2014. The adoption of EHRs is being pushed with the belief that the information contained in EHRs will improve medical decision making with an associated improvement in patient outcomes [2].

Information visualization systems have been developed to facilitate the synthesis and analysis of large amounts of information using temporal and sequence analysis [7]. This project demonstrates a visual analytic tool that grew organically from a question and collaboration between a physician and computer science engineers. This tool is designed to address specifically challenges to the extraction of meaningful information from EHR data. We developed a time-stamped information visualization tool, VisCareTrails, to facilitate the analysis of patient histories stored in the EHR. The use case will use VisCareTrails to focus on the diagnosis of pancreatic cancer.

VisCareTrails is a system based on timed word tree visualizations summarizing event paths relative to a given root event. These are generated in a simple drag-and-drop user interface. In particular, in this domain of patient histories, we see VisCareTrails as an interesting alternative to a previous visualization called LifeFlow [6]. This process summarizes multiple sequences of timed-events and generalizes the idea of Word Trees [8]. VisCareTrails provides the user a means to explore electronic health data in order to understand patterns, problems and opportunities in clinical practice.

- *Lauro Lins is with NYU-Poly, E-mail: lauro@nyu.edu.*
- *Marta Heilbrun is with Departament of Radiology, Univ. of Utah, E-mail: marta.heilbrun@hsc.utah.edu.*
- *Juliana Freire is with NYU-Poly, E-mail:juliana.freire@nyu.edu.*
- *Claudio Silva is with NYU-Poly, E-mail: csilva@nyu.edu.*

2 A Visual Summary for Event Sequences

The central element of VisCareTrails is a visualization that summarizes multiple event sequences. The idea is to summarize S, an input set of event sequences, based on another input: a root event, r. Once S and r are defined, a visual summary is generated in two steps. First, an *event tree*, T, based on S and r is computed. Second, a visual representation, V, for the event tree, T, is generated.

2.1 Event Trees

An *event tree* is a simple way to summarize event sequences. Figure 2 shows an example of such an object. Given event sequences S and a root event r, the first step is to choose an *alignment point* for each input sequence. In Figure 2, alignment points are indicated by red circles and i_r define their indices. The event at the alignment point of each sequence should be equal to the root event (C in our example). Once the alignment points are defined, we add a root node to the event tree with label r, offset 0, and set all sequences from S as members of this node (*e.g.,* central node of T in Figure 2). Next, a left parse (negative offsets) and right parse (positive offsets) on each input sequence starting from its alignment point is performed. In our example, the left parse of s_1 generates first the node with label B and offset -1 and then the node with label A and offset -2 (note that s_1 is present in these two nodes). The right parse of s_1 generates first the node with label D and offset 1, and then the node with label E and offset 2, both having sequence s_1 as a member. We follow the same idea for the left and right parses of the other sequences always reusing existing nodes when possible. For example, when doing the left parse of s_3, we reuse the same node with label B and offset -1 as the one generated when left parsing s_1.

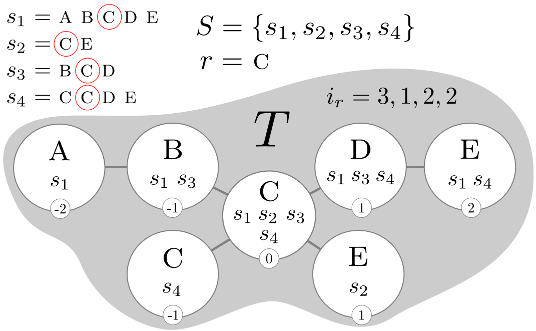


Fig. 2. Example of an event tree, T, rooted at event C for the set of event sequences S. Each node in T has an event label, a subset of sequences, and an offset (small circle). The central visualization in VisCareTrails are visual representations for event trees.

2.2 LifeFlow

The concept of event trees has been shown useful for the problem of making sense of patient histories. Wang et al. [6] define a sentinel event (our root event) as a way to align temporal data and find patterns once the alignment is established. Later, Wongsuphasawat et al. [9] proposed LifeFlow, a technique that computes an event tree T and then generates a visual encoding for it: V_{LF}. Figure 3 shows a LifeFlow visualization for a dataset of hospital events regarding arrivals, transfer between blocks (ICU, Emergency, Floor), discharges, and deaths. In V_{LF}, the nodes of T are graphically encoded by rectangles and their labels are encoded by colors (a legend is necessary to map colors into event names). The height of each rectangle's node is proportional to the number of sequences in its node and the width is proportional to a summary measure (*e.g.,* mean) of the time difference between the node's event and the previous event for all its sequences. The left side of a child node's rectangle intersects completely the right side of the rectangle of its parent node.

Although we considered using the LifeFlow visual summary as the central display in VisCareTrails, two problems drove us to a different visualization. First, the datasets we plan to analyze with VisCareTrails contain thousands of event types (*e.g.,* diagnostic exam names). It is unfeasible to associate a fixed color to each event type and let a user learn this association once. To understand event paths with LifeFlow in our use case, a continuous back-and-forth effort between the main visualization and the color translation legend is required. The second problem is that we want to support dozens of simultaneous event types in a single visualization. In this case, even with the color translation legend, it is hard to read the main LifeFlow visualization, because it is hard to perceive different colors when more than just a few colors (*i.e.,* less than a dozen) are used.

2.3 Timed Word Trees

Inspired by *Word Tree* displays [8], our basic idea was to replace colored rectangle labels used in LifeFlow visualizations with text labels. If this could be done while preserving, to a certain degree, the other characteristics of LifeFlow visualizations, we would obtain a better central visualization for VisCareTrails (*e.g.,* without the two problems mentioned before).

Why not standard word trees? In fact word trees is an interesting alternative to visually encode paths and path frequencies for an event tree. The problem is that one piece of information present in event trees and encoded in LifeFlow visualizations is not encoded in a standard word tree: the time distance between two events (two adjacent nodes in an event tree). To address this issue we propose *timed word trees*, a generalization of word trees where each word in the tree has an associated time stamp and the final display encodes the time distances between the words based on these time stamps. Figure 1 shows a timed word tree for pancreatic cancer patients. From this display we can read that the average time span between the last stage event and the death event decreases for patients that die when officially registered in, respectively, Stage I, Stage II, Stage III and Stage IV. A more elaborate timed word tree example is shown in Figure 4 (same event tree as Figure 4).

Equally spaced guide-lines are rendered in order to help convey the concept of time on a timed word tree. One of the characteristics of a timed word tree is that, although time order is preserved, equal display lengths might represent different time lengths. To help minimize this distortion, we map the guide lines crossing the visualization back into a linear time line (see the the light green, gray, blue transition rectangles on the timed word tree displays). Note, for example, that the guide-lines that cross the Dead node in Figure 1 are all mapped to the same point on the light blue rectangle.

Our current algorithm to render timed word trees involves (1) opening space in the time axis to fit event label dimensions and time distances, and (2) setting a y coordinate to the words (assuming x coordinate is time) so as to avoid text overlaps and at the same time have a packed layout. A detailed explanation of (1) and (2) is beyond the scope of this paper, but it is worth mentioning that the algorithm is fast: $O(n\log(n))$, where n is the number of words, and we are able to layout timed word trees with millions of nodes in a fraction of a second using a standard laptop.

3 VisCareTrails System

VisCareTrails supports the following pipeline: (1) a set of time-stamped event sequences is loaded into the system; (2) *group-events* are defined as needed (Stage III in Figure 1 is a group-event that means either event III, IIIA, IIIB or IIIC); (3) a timed word tree is generated by dragging and dropping events and/or group-events into the central canvas (in Figure 1, stage events & Dead were dragged and dropped into the canvas); (4) one of the dropped events is defined as the root event (by default the root is the first element that was dropped in the visualization, but a user can change the root event at any time); (5) the visual summary generated is inspected to understand paths that end and start in the root event; and (6) path nodes are selected to obtain survival curves for the sequences. Figure 1 shows survival curves of the selected stage nodes (red, green, purple, and orange paths): bottom left widget. The visual summary conveys information about frequency of events (larger fonts and thicker transitions means more sequences going through the path), time distances (based on average times) of the events relative to their parent event; and a hint on the dispersion

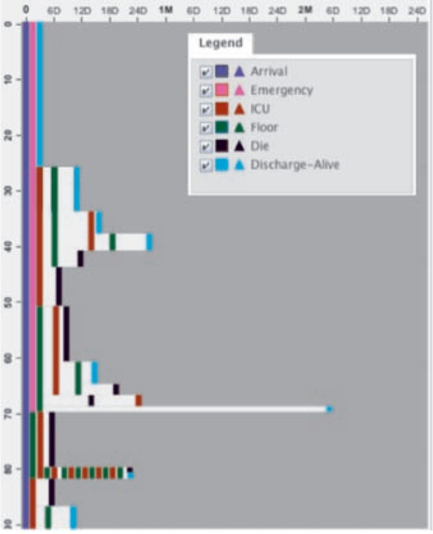


Fig. 3. LifeFlow visualization, V_{LF}, summarizing hospital event sequences for 91 patients (taken from [9]).

(*i.e.,* standard deviation) of time distances in each event transition (i.e. the hue of blue darkens as the standard deviation of the time distance decreases). On the second bottom widget (from left to right), we show a box-plot for the time distance distribution from the selected events to the root event.

4 Pancreas Cancer Use Case

4.1 Pancreas Cancer

Cancer represents a unique case in which to use EHR data to study health care complexity. Cancer of the exocrine pancreas is the fourth leading cause of cancer death in the US. In 2010, it was estimated that 43,140 new cases and 36,800 deaths occurred from pancreatic cancer in the US, with only 6% overall survival at 5 years [1].

4.2 Patient Cohort

Since 2000 more than 1300 cases of Cancer of the Pancreas have been diagnosed in the State of Utah. Many of these patients are triaged to a single National Comprehensive Cancer Network tertiary care cancer center. This center maintains cancer patient data in an electronic data warehouse. The pancreatic cancer patient data on 631 subjects was extracted in the summer of 2010. In this initial pass, 17,780 unique events, recorded from an EHR, including cancer stage details, vital status, radiology and other diagnostic procedure codes, and laboratory tests were imported into VisCareTrails. In order to comply with patient privacy rules, the event data was extracted from a data warehouse, and the subjects were anonymized.

4.3 Cancer Survival

This use case demonstration of VisCareTrails establishes that the information in the EHR can be read into the visualization program, and that the record of events is intuitively accurate.

The VisCareTrails display in Figure 1 demonstrates an expected distribution of patients and expected outcomes. According to the American Cancer Society, the five year survival for local and regional disease is 31%, while less than 20% of patients present with low enough stage disease to be considered surgical candidates [1]. In our population, the tree intuitively and quantitatively demonstrates the survival. Two-thirds (64%) of subjects present with advanced stage disease. The median survival for the 9% of the population who presents with Stage I disease is almost 750 days but only 200 days for subjects who present at Stage IV. This visual information mirrors that which is generated statistically by a Kaplan-Meier survival curve, however is intuitive to the physician end-user, and bypasses interaction with a statistical program.

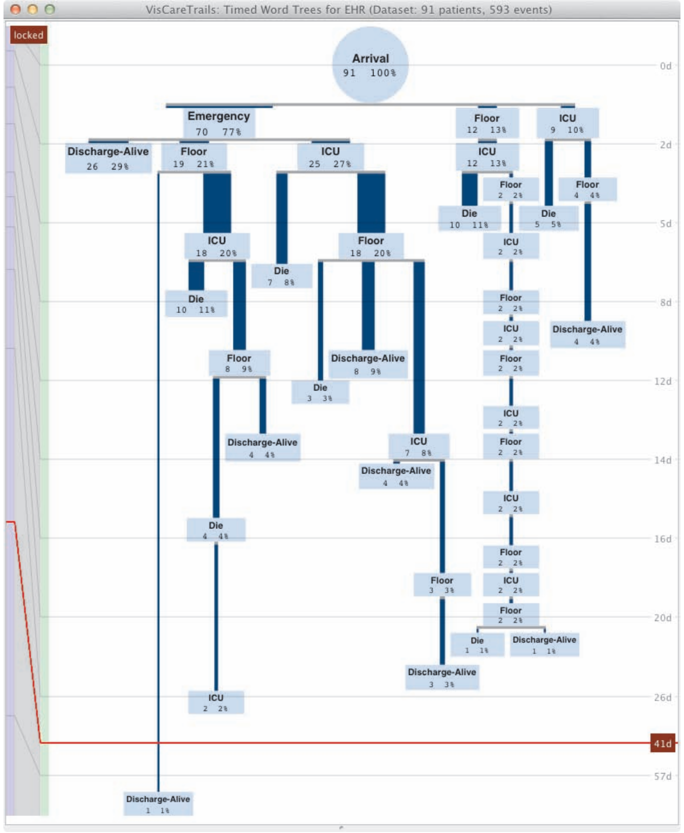
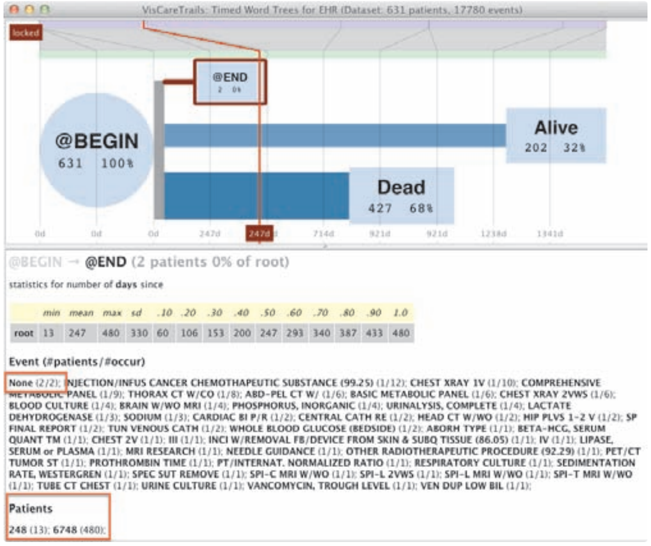


Fig. 4. Proposed timed word tree visualization, V_{TWT}, in VisCareTrails for the same event tree of Figure 3.

4.4 Identification of unclean and missing data

In the database, the records of Dead ($n = 427$) or Alive ($n = 202$) are recorded. For two patients an assessment of vital status is unknown (Figure 5). Four of the subjects had events that took place after the Dead event. When the tree is rooted on Dead, these events appear as positive branches. This type of unclean data is easily identified in the visualization tool.



@BEGIN → @END (2 patients 0% of root)

statistics for number of days since

	min	mean	max	sd	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.0
root	13	247	480	330	60	106	153	200	247	293	340	387	433	480

Event (#patients/#occur)

None (2/2); INJECTION/INFUS CANCER CHEMOTHAPEUTIC SUBSTANCE (99.25) (1/12); CHEST XRAY 1V (1/10); COMPREHENSIVE METABOLIC PANEL (1/9); THORAX CT W/CO (1/8); ABD-PEL CT W/ (1/6); BASIC METABOLIC PANEL (1/6); CHEST XRAY 2VWS (1/6); BLOOD CULTURE (1/4); BRAIN W/WO MRI (1/4); PHOSPHORUS, INORGANIC (1/4); URINALYSIS, COMPLETE (1/4); LACTATE DEHYDROGENASE (1/3); SODIUM (1/3); CARDIAC BI P/R (1/2); CENTRAL CATH RE (1/2); HEAD CT W/WO (1/2); HIP PLVS 1-2 V (1/2); SP FINAL REPORT (1/2); TUN VENOUS CATH (1/2); WHOLE BLOOD GLUCOSE (BEDSIDE) (1/2); ABORH TYPE (1/1); BETA-HCG, SERUM QUANT TM (1/1); CHEST 2V (1/1); III (1/1); INCI W/REMOVAL FB/DEVICE FROM SKIN & SUBQ TISSUE (86.05) (1/1); IV (1/1); LIPASE, SERUM or PLASMA (1/1); MRI RESEARCH (1/1); NEEDLE GUIDANCE (1/1); OTHER RADIOTHERAPEUTIC PROCEDURE (92.29) (1/1); PET/CT TUMOR ST (1/1); PROTHROMBIN TIME (1/1); PT/INTERNAT. NORMALIZED RATIO (1/1); RESPIRATORY CULTURE (1/1); SEDIMENTATION RATE, WESTERGREN (1/1); SPEC SUT REMOVE (1/1); SPI-C MRI W/WO (1/1); SPI-L 2VWS (1/1); SPI-L MRI W/WO (1/1); SPI-T MRI W/WO (1/1); TUBE CT CHEST (1/1); URINE CULTURE (1/1); VANCOMYCIN, TROUGH LEVEL (1/1); VEN DUP LOW BIL (1/1);

Patients

248 (13); 6748 (480);

Fig. 5. Data cleaning: by dropping @BEGIN, Dead, Alive and @END events we are able to visually identify a path that shouldn't exist: from @BEGIN direct to @END. Mouse hovering on this path we get a report showing two patient identifiers and their events between @BEGIN and @BEGIN. Highlighted None event also requires further investigation.

4.5 Detection of diagnostic testing strategies

The most commonly utilized diagnostic test in the cohort is a CT of the abdomen and pelvis CT_AP, of which 424 patients underwent a total of 1469 examinations. Figure 6 shows the most common sequences of diagnostic tests in the Stage IV group of patients. This interface readily demonstrates the types, frequencies and sequences of tests that occur in the cohort. The hypothesis that prompted this visualization tool is that differences in survival can be attributed to different diagnostic tests. An evaluation of Surveillance Epidemiology and End Results-Medicare-linked data from 2010 [4] suggested that patients with pancreatic cancer who underwent an endoscopic ultrasound (EUS) had improved survival compared to those who did not. We attempted to replicate this analysis in our data, by looking at subjects who underwent EUS. However there were only 48 such subjects in the cohort, making any analysis limited because the absolute number of events per node tended to be very small.

5 Discussion

5.1 Data interpretation and domain expertise

The interaction and impact of a domain expert in the design of this tool is an essential component of the tool development. The hypothesis that prompted this visualization tool is that differences in survival can be attributed to different diagnostic tests. In one pass, examining the utilization of PET, a curve was generated showing that subjects who had a PET < 70 days after the staging event had a shorter survival than those who had a PET > 70 days after the staging event (not shown). This might suggest that an early PET was associated with poorer outcomes. However, the physician, suggested rather, that the subjects who were alive > 70 days after staging, just by being alive, had more opportunities for surveillance imaging.

In regards to the question of the role of the EUS in the diagnosis, the domain expert deemed 48 an unrealistically low number. The information brought into the tool only pulled from the primary diagnosis procedure codes (ICD code). Because multiple procedures may be coded in a single setting, that is to say an endoscopic retrograde cholangiopancreatogram (ERCP) will be performed in the same setting as an EUS; we may have caught the primary code for the ERCP, but missed the secondary procedure code for the EUS. It will be necessary to pull the secondary procedure codes into the database in order to run this analysis.

Heterogeneous information will be a part of any EHR and subsequent analysis as the uptake of these records is inconsistent, and data standards do not yet exist [3]. The interaction between physicians and clinical experts and the systems that make it a simple process to identify of the data that is missing, unavailable, or in error is essential to optimize the analysis process of the EHR. Some of the inefficiencies in medicine may be due to events that do not occur and should, such as recommended screening [5]. Visualization tools may facilitate the process of identifying steps not taken.

5.2 Limitation

Timed word tree visualizations require events to follow the exact order in which they happened. This is useful for creating a snapshot of the events that lead up to the end of the study period or death. However, it may be that it is not exactly the sequence of events or tests but rather the specific combination of events or tests that segregate populations. Until it is possible to create distinct groups of test populations (*e.g.,* patients who had CT_AP and EUS compared to patients who had CT_AP, EUS and MRI, regardless of whether the EUS or the CT_AP was the first event) we may be missing relevant patterns in the data.

6 Conclusions

Time stamped information visualization tools, like VisCareTrails, capture EHR patient events and display the information in an intuitive fashion. This makes it very useful for the purposes of analyzing a record when there is a discrete start and end event, such as cancer records. However, challenges persist in optimizing the tool to tease out both diagnostic testing strategies and bundled events that are associated with differences in survival.

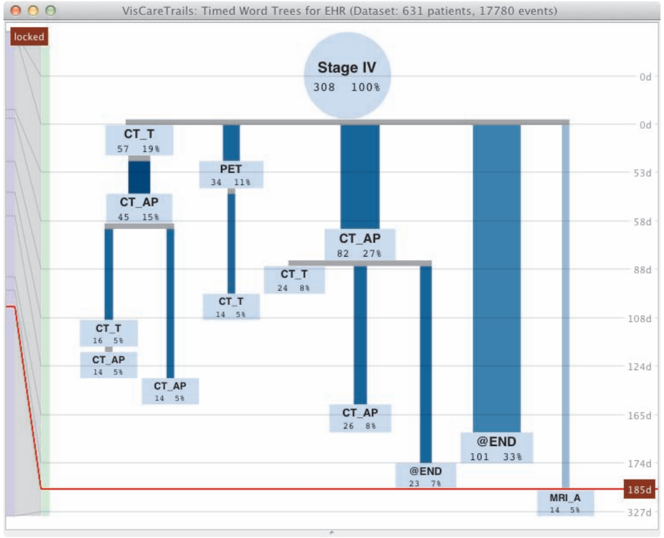


Fig. 6. Timed word tree with most frequent event paths (≥ 14 patients) after a patient gets registered in Stage IV. Events considered are in diagnostic test groups CT_AP, CT_T, EUS/ERCP, MRI_A, or PET. Event @END was included to indicate frequent paths where no event in a diagnostic test group occurred (*e.g.,* 33% of the patients are not tested in any of the considered diagnostic tests: thick branch leaving root event). Branches are sorted by average transition time.

Acknowledgments

The authors wish to thank Isaac Francis, Samir Courdy and Joslyn Christensen from the Huntsman Cancer Institute. This work was supported in part by the NCCN and CTSA grants to the University of Utah.

References

[1] American cancer society: Cancer facts & figures 2010. Technical report, American Cancer Society, Atlanta, 2010.

[2] D. Blumenthal. Promoting use of health it: why be a meaningful user? *Maryland medicine*, 11(3):18, 2010.

[3] K. Hayrinen, K. Saranto, and P. Nykanen. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5):291–304, may 2008.

[4] S. Ngamruengphong, F. Li, Y. Zhou, A. Chak, G. S. Cooper, and A. Das. Eus and survival in patients with pancreatic cancer: a population-based study. *Gastrointestinal endoscopy*, 72(1):78–83, 83 e1–2, Jul 2010.

[5] H. Singh, K. Hirani, H. Kadiyala, O. Rudomiotov, T. Davis, M. Khan, and T. Wahls. Characteristics and predictors of missed opportunities in lung cancer diagnosis: An electronic health record–based study. *Journal of Clinical Oncology*, 28(20):3307, 2010.

[6] T. Wang, C. Plaisant, A. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman. Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 457–466. ACM, 2008.

[7] T. Wang, K. Wongsuphasawat, C. Plaisant, and B. Shneiderman. Extracting insights from electronic health records: Case studies, a visual analytics process model, and design recommendations. *Journal of Medical Systems*, pages 1–18, 2011.

[8] M. Wattenberg and F. Viégas. The word tree, an interactive visual concordance. *— IEEE Transactions on Visualization and Computer Graphics*, pages 1221–1228, 2008.

[9] K. Wongsuphasawat, J. Guerra Gómez, C. Plaisant, T. Wang, M. Taieb-Maimon, and B. Shneiderman. Lifeflow: visualizing an overview of event sequences. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1747–1756. ACM, 2011.

AnamneVis: A Framework for the Visualization of Patient History and Medical Diagnostics Chains

Zhiyuan Zhang[1], Faisal Ahmed[1], Arunesh Mittal[1], IV Ramakrishnan[1], Rong Zhao[1], Asa Viccellio[2], and Klaus Mueller[1]

[1]Computer Science Department and Center for Wireless and Information Technology (CEWIT)

[2]Department of Emergency Medicine

Stony Brook University

Abstract

The medical history or *anamnesis* of a patient is the factual information obtained by a physician for the medical diagnostics of a patient. This information includes current symptoms, history of present illness, previous treatments, available data, current medications, past history, family history, and others. Based on this information the physician follows through a medical diagnostics chain that includes requests for further data, diagnosis, treatment, follow-up, and eventually a report of treatment outcome. Patients often have rather complex medical histories, and visualization and visual analytics can offer large benefits for the navigation and reasoning with this information. Here we present *AnamneVis*, a system where the patient is represented as a radial sunburst visualization that captures all health conditions of the past and present to serve as a quick overview to the interrogating physician. The patient's body is represented as a stylized body map that can be zoomed into for further anatomical detail. On the other hand, the reasoning chain is represented as a multi-stage flow chart, composed of date, symptom, data, diagnosis, treatment, and outcome.

Keywords: health care, medical record presentation, EHR, EMR

1 Introduction

The electronic health record (EHR) digitally stores patient health information generated by one or more clinical encounters in any care delivery setting. This information includes patient demographics, problems, symptoms, diagnoses, progress notes, treatments, medication, vital signs, past medical history, immunizations, laboratory data, radiology reports, and many others. However, the acceptance of the EHR in clinical practice lags far behind its expectation and potential. Related information and overviews are typically difficult to obtain, severely impeding a physician's diagnostic reasoning. The inefficient, fragmented display of patient information is a likely cause. In this paper we offer a first step to overcome these deficiencies by comprehensibly organizing the patient medical history, also known as *anamnesis*. We employ the concept of *Five W's* (*who*, *when*, *what*, *where*, *wh*y, and also *how*) of journalistic reporting to structure the medical information domain and provide a suitable visual mapping for each for visual information display.

The Five W's are the elements of information needed to get a full story. They are encountered in many playing fields: by a journalist uncovering a political scandal, a police detective investigating a crime, a customer service representative trying to resolve a complaint, and a market analyst planning an effective marketing campaign. The order in which the information is gathered or interrogated can vary case by case – crucial is only that all five W's are ultimately addressed.

When it comes to applying the Five W's to visualization design, we can break it down into two steps: (1) identify all Five W components and their relations, and (2) map these to suitable visual information encodings and interactions.

We propose to use the Five W's in our health care informatics application as a means to establish a comprehensive multi-faceted assessment of the patient and his (her) history for intuitive information retrieval. The goal is information organization and integration along these various aspects. Overview and detail-on-demand requires hierarchies, and effective information organization requires robust encoding by ways of well-established criteria – we use standard codes commonly used for diagnosis and billing in hospitals which enables us to easily build our system on top of an existing health care information system. These codes are ICD, CPT, and NDC. ICD is the code used to describe the condition or disease being treated, also known as the diagnosis. CPT is the code used to describe medical services and procedures performed by doctors for a particular diagnosis. NDC is the code used for administered drugs. ICD is widely accessible (developed by the World Health Organization), CPT is proprietary and only available to healthcare providers, and NDC is also publicly available. Further goals, often expressed by our collaborating emergency physician – who is also a co-author of this article – are ease of information access and flexibility in displayed aggregated information and data. To enable this functionality, our system is fully interactive and the displays are fully linked and coordinated.

2 Related Work

A number of approaches for the visualization of medical patient records have been proposed, and new systems are likely to emerge as the Electronic Health Record (EHR) is adopted widely. A frequent paradigm is to organize the patient records along the time axis. Prominent efforts in that direction are LifeLines [7] and LifeLines2 [11] in which health records are distinguished by their inherent aspects, such as problems, symptoms, tests/results, diagnosis, treatments and medications, etc. and color is used to indicate severity or type A level of detail mechanism allows one to zoom into patient records. A number of other works, such as [6], have also embraced this type of patient data visualization. Particularly interesting in this context is the work of Aigner et al. [2] who have made use of illustrative abstractions to gradually transition between broad qualitative overviews of temporal data (for example, blood pressure) to detailed, quantitative time signals. These techniques are part of the Midgaard system [3] which also provides a visualization scheme in which acquired patient data are mapped to a template of a human body (although little further detail on how this scheme is used in practice is available). The system described in [8] gathers close-ups of acquired radiological data around a volume-rendered full body. In fact, many modern EHR systems now support time-line views and are also beginning to support body-centric data layouts. Another frequently used paradigm is that of flow-charts, as used in clinical

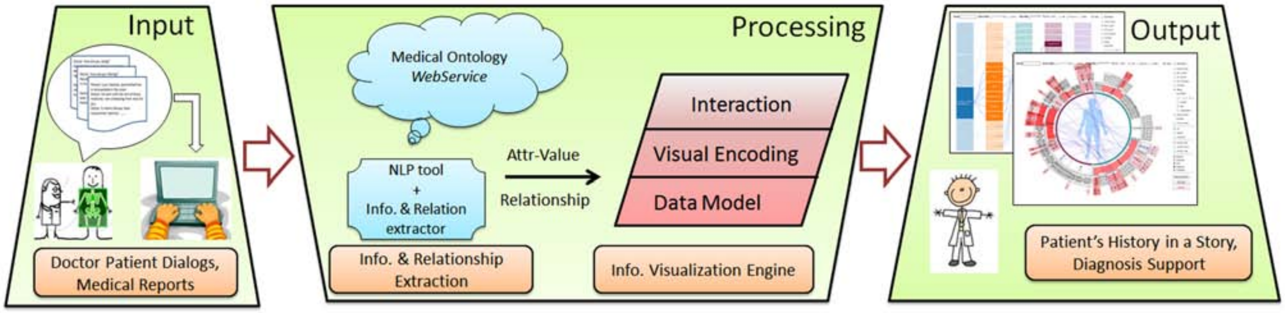

Input
Doctor Patient Dialogs, Medical Reports
Processing
Medical Ontology WebService
NLP tool + Info. & Relation extractor
Attr-Value Relationship
Interaction
Visual Encoding
Data Model
Info. & Relationship Extraction
Info. Visualization Engine
Output
Patient's History in a Story, Diagnosis Support

Figure 1: System Pipeline

algorithm maps [5] and others [4][10], where patient records are visualized as a logical execution sequence of plans. These methods typically operate without temporal alignments. Finally, works also exist that combine these two paradigms into coordinated views [1].

While our recent work [12] also embraced the Five W's scheme, its main focus was a visual interface that a doctor might use to log and review evidence gathered (and actions required) during a patient-doctor dialog (called *encounter*). This system combined the temporal functionality of [7][11] with the body-centric data arrangement of [3][8] and supported analytical reasoning with these information items via a force-directed graph (called the *diagnosis sandbox*).

When cast into the Five W's we find that most existing systems support the *when*, *what*, *why*, and *where* aspects quite well, although few support all of these. Functionality for coordinated views linking specialized visualization for these aspects is less supported. Apart from this, a further main difference to existing systems is our representation of the *who*. While most systems reduce it to simple personal data, such as name, age, gender, smoker, and the like, we see it as an opportunity to represent all medical information ever recorded about a patient – a true reflection of the person (in terms of medical history at least). All is captured within a modern information visualization framework and linked with the other coordinated displays for the other 4 W aspects.

3 IDENTIFYING THE FIVE W'S: INFORMATION EXTRACTION

The information flow of our system is summarized in Fig. 1. The input to our system are patient records and medical reports, doctor-patient dialogs and other interactive inputs, results from triage, and data acquired from the patient, such as radiological images, lab analyses, and the like. At the processing stage an NLP (Natural Language Processing) engine cooperates with an online medical ontology server to extract structured information and relationships from this incoming information and data stream. It then formats the extraction results into the Five W model and passes it on to the visualization engine. The visualization engine has all procedures and data models to encode the Five W information facets into corresponding visuals and interaction procedures. The output of this process is then presented in the visual interface that is subject of this paper. In the box labelled 'Output' we show two displays: (i) a hierarchical radial ring display (foreground window) that visualizes the patient history in the context of a centered body map and (ii) a sequential (causal) display (background window) that visualizes the diagnostic reasoning chain. Before we describe this interface in detail, we first discuss the conceptual information organization of our system, in terms of the structuring Five W's.

3.1 The *Who* and *What*

The *who* and *what* information helps doctors to quickly assess the history and status of the patient. It describes the patient in terms of:

♦ Symptoms and Diagnosis: this includes the patient's symptoms, injuries, and any diagnosed diseases. All of this information can be encoded using the ICD code standard.

♦ Procedures: these include patient tests and examinations, treatments administered, and drugs prescribed. This type of information can be encoded using the CPT code or the ICD-procedure code standard, and the NCD code standard.

♦ Data: these include test and examination results, review of systems, vital signs, and social and family history. The codes for these are part of the procedure code and yields information on what the patient already has.

Our system encodes this information in two ways: in a hierarchical radial ring display and in a sequential (causal) display.

3.2 The *Where*

The *where* information refers to the location of the *who* and *what* information within the confines of the patient's body. While not all information can be localized that way, for the information that can be localized, we encode it in a body outline map surrounded by the ring display. Items on the ring display are pointing to the appropriate locations on the body. The Google Body Browser [13] could then be indexed by a subset of the *what* and so give the doctors a good start for further exploration and also offer explanations to the patient.

3.3 The *When*, *Why,* and *How*

The *when*, *why,* and *how* show a case under (doctor) collaborative diagnosis/treatment, or an entire life span. It demonstrates for each node *what, when, why*, and *how* that node appears. Various multi-resolution and selection techniques are available to make the visualization scalable. It supports two types of displays: a sequential display and a hierarchical radial display.

The sequential display stresses causal relationships and encourages causal reasoning done by the doctor. It also aims to model the typical medical workflow: (1) observe symptoms and possibly browse history data, (2) prescribe and evaluate tests results, (3) form hypotheses and possibly acquire more data, (4) cast diagnoses and (5) prescribe treatments. These steps may all be executed within one patient visit or they may prolong over some period of time, but the overall workflow is always engaged. The 5th step may include a referral to another doctor, which then starts another workflow (back-linking to the previous).

The hierarchical radial display aims to provide an overview of the entire history of the patient, offering detail on demand.

4 Encoding the Five W's: Information Visualization

We have two types of cooperating displays:

♦ A hierarchical radial (patient overview) display with an integrated body outline primarily for the *who* and *where.*

♦ A sequential (diagnostic reasoning) display primarily for the *when*, *why*, and *how.*

The *what* is part of both displays (in form of the various nodes) and is context-sensitive. The two interfaces are linked, such that operations on either view will be reflected in the other. Thus, one can quickly switch between the sequential (and possibly evolving) diagnostic reason flow and the radial patient overview display.

4.1 Hierarchical Radial Display – The 'Who' Display

The hierarchical radial display is used primarily to show the *who* and *where* information of the patient. Based on the category information discussed in Section 3.1, the *who* includes three radial displays, one for symptoms and diagnoses, one for procedures and treatments, and one for data. These three displays are interlinked to allow doctors to obtain a full picture of the patient as well as assess existing relationships.

4.1.1 Data Model

We use a tree data structure to store the code hierarchy information. For each symptom or diagnosis the patient has, we find the node *n* in the tree with the corresponding ICD9 (soon ICD10) code, and insert the new item as a child for node *n*. For example, if the patient has *bacterial meningitis* whose ICD9 code is 320, we first build an incident node *m* for this diagnosis to store its information (severity, result, etc.). In the tree, we find the node *n* with code 320, which is [320 *bacterial meningitis*]. Then we insert *m* as a child of *n*. After this, we update all ancestors of *n* with the new inserted incident node's information, such as number of incidents that fall into this category, severity, and so on. By doing this for all symptoms, diagnosis, and procedures, the tree will always be current and contain the patient's entire history.

4.1.2 Visual Design

We use the *sunburst* visualization paradigm [9] (see Fig. 3) to visually convey the tree structure. A sunburst is a polar-coordinates hierarchical space-filling diagram. Nodes in the sunburst layout are drawn as solid areas (either wedges or bars), and their placement relative to adjacent nodes reveals the relationships in the hierarchy. Because the nodes are space-filling the angle for each node can be used to encode additional information, such as number of incidents in our case.

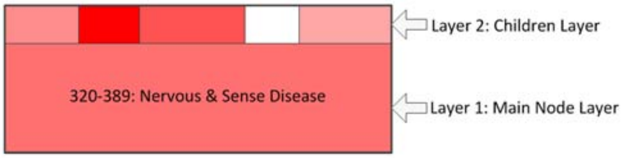


Figure 2: Node design. Color encodes severity. The main node layer tells us that the patient has a relative severe disease in the nervous and sense organs. The children layers provide more detail with regards to what the diseases are.

Each node has a wedged shape in the sunburst tree. We further decompose the node into three layers to encode more information, as is shown in Fig. 1. These layers are:

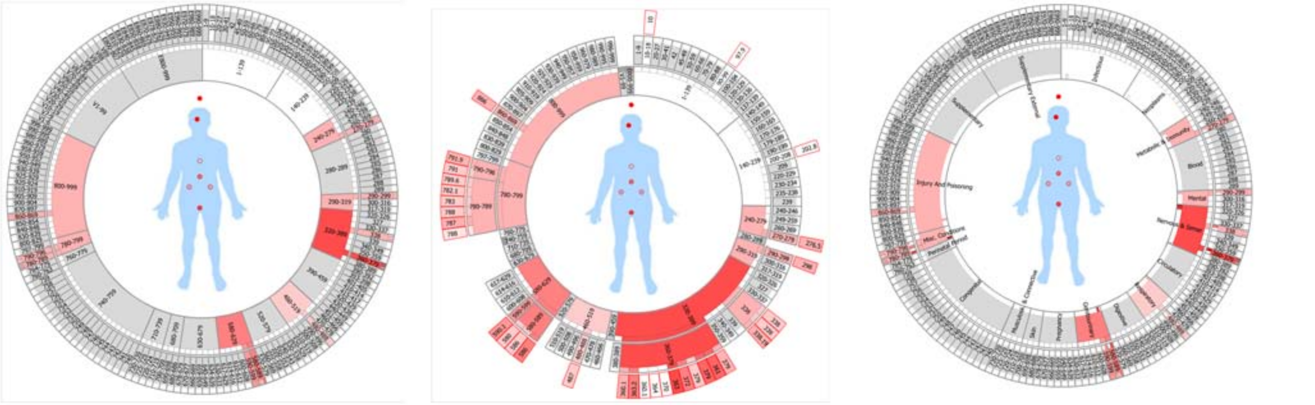
♦ *Layer 1:* the main node layer which is used to display information about the node, such as codes, name, etc.

♦ *Layer 2:* encodes the next lower level in the hierarchy. It is meant to give users a quick overview on the sub-diseases without showing their real nodes. We provide this layer to make the hierarchy display scalable.

Each radial display is either hierarchy-centric or patient-centric. In the hierarchy-centric display (Fig. 3a), each node in the sunburst tree is sized by how many sub-categories it has. It focuses more on the hierarchy information represented in the medical codes and serves as an illustration of the complexity of a sub-system and its composition. In the patient-centric display on the other hand, more radial space is dedicated for diagnoses/procedures the patient had activities in. For categories that the patient does not have any activities in, the node will be collapsed to save space for others (see Fig. 3b).

There are three levels of code hierarchies in the sunburst radial display. The first level corresponds to the highest code hierarchy level. The second level shows more detailed categories. The third level contains the incident nodes, which are the medical items (symptoms/procedures/diagnosis) that the patient has activities in. Three default level filters are provided to help users quickly explore these three levels. Also users can expand and collapse the nodes interactively by their expertise.

The root of the tree is displayed in the center of the sunburst. However given the sole application context – the patient – we chose to replace the standard root node by a body outline. This enables us to intuitively fuse the *who* with the *where* display. If an incident (medical record) has corresponding location information,



(a) Hierarchy-centric layout

(b) Patient-centric layout

(c) Display with ICD code name/label overlay

Figure 3: Sunburst display for symptoms and diagnoses

a red dot is displayed in the body outline. The intensity is used to encode the severity. Thus by looking at the body outline, doctors can quickly learn which parts of the patient's body have (or had) diseases and also judge their severity by the color intensity. Hovering on the red dots will popup more details about the injured part, such as name, severity, and how many incidents are related. Clicking the red dot will highlight the corresponding diseases in the sunburst tree.

Finally, users unfamiliar with the ICD9 coding system have the option to display the ICD node names and labels as an overlay (see Fig. 3c). Since the nodes are always in the same relative positions in the sunburst display, experienced users may soon acquire a mental map of the system and only require the overlay in non-routine situations.

4.2 Sequential (Causal) Display

The sequential display (Fig. 4) is used mainly to demonstrate the *what*, *when*, *why* and *how* information, which embodies the medical diagnostic flow. The medical records are organized by an underlying graph data structure. Each node corresponds to one incident (medical primitive), which could be a doctor visit, symptom, test/data, diagnosis or treatment. Edges represent relationships.

4.2.1 Visual Design

A node is displayed as one elongated box because it better utilizes the rectangular screen, better fits the text, and has better scalability compared to a circular shape. All of our medical collaborators agreed on this. If two nodes are related with one another, an edge is drawn to link them together. Edge bundling is used to reduce cluttering. Usually the diagnostic workflow is: Patient visits doctor → patient complains about symptoms → doctor orders tests for patient → doctor renders a diagnosis → treatments are given → outcome is observed. Thus, the sequential display can show these reasoning chains very well. In some cases the current doctor refers the patient to see another specialist (which is the treatment in this case), or current symptoms are caused by previous described drugs (which can be a form of diagnosis). In situations back edges appear. Back edges are shown in different color (red) to make them easy to see. Back edges may be due to treatments causing new symptoms, or they may be treatments constituting doctor referrals. Fig. 4 has no back edges.

5 Conclusions

We have presented an application of the Five W's scheme of information gathering and reporting, with a special application to health care informatics. While our informal user studies are highly encouraging and promising, we would like to conduct more formal studies next. We have collaborations with six leading physicians at our university, ranging from radiology, neurosurgery, family medicine, and emergency medicine. Using well-defined tasks, we would now like to test our system with these individuals and also with the good-sized population of medical students they educate. Finally, a second possible user group for our framework are medical coding personnel who work in the hospital billing office to translate medical records to ICD9 code. Our sunburst display has good potential for them to better recognize relationships in medical services and so perform more accurate billing statements. We are currently pursuing efforts on this level as well.

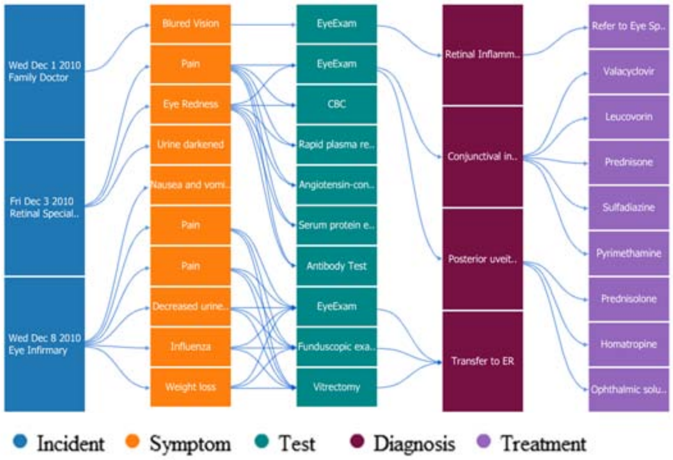


Figure 4. Sequential display for diagnostic chain.

Acknowledgments

We would like to express our deepest thanks to the UC Berkeley Visualization Lab and Stanford HCI group to make available the very helpful Flare toolkit, which we used as a basis and extended non-trivially for the system presented in this paper.

References

[1] W. Aigner, S. Miksch, "Supporting Protocol-Based Care in Medicine via Multiple Coordinated Views," *Proc. Coordinated and Multiple Views in Exploratory Visualization*, pp. 118-129, 2004.

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

[3] R. Bade, S. Schlechtweg, S. Miksch, "Connecting Time-oriented Data and Information to a Coherent Interactive Visualization," *Proc. Human Factors in Computing Systems (CHI),* pp. 105-112, 2004.

[4] J. Fox and R. Thomson., "Decision Support and Disease Management: A Logic Engineering Approach," *IEEE Transactions on Information Technology in Biomedicine*, 2(4):217–228, 1998.

[5] D. Hadorn, "Use of Algorithms in Clinical Practice Guideline Development: Methodology Perspectives," *AHCPR Pub.*, 0009(95):93–104, Jan. 1995.

[6] R. Kosara, S. Miksch, "Visualization Techniques for Time-Oriented, Skeletal Plans in Medical Therapy Planning," *Proc. Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making (AIMDM),* pp. 291-300, 1999.

[7] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, B. Shneiderman, "Lifelines: Using visualization to enhance navigation and analysis of patient records," *Proc. AMIA Annual Symposium*, pp. 76–80, 1998.

[8] T. Ropinski, I. Viola, M. Bierman, H. Hauser, K. Hinrichs, "Multimodal Visualization with Interactive Closeups," EGUK Theory and Practice of Computer Graphics (TPCG), 2009.

[9] J. Stasko and E. Zhang. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. *IEEE Symposium on Information Visualization, 2000. InfoVis 2000*, pages 57-65, 2000.

[10] S. Quaglini, M. Stefanelli, G. Lanzola, V. Caporusso, S. Panzarasa, "Flexible guideline-based patient careflow systems," *Artificial Intelligence in Medicine*, 22(1):65–80, 2001.

[11] T. Wang, C. Plaisant, A. Quinn, R. Stanchak, B. Shneiderman, and S. Murphy, "Aligning temporal data by sentinel events: Discovering patterns in electronic health records," *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 457-466, 2008.

[12] Z. Zhang, A. Mittal, S. Garg, A.Dimitriyadi, IV Ramakrishnan, R. Zhao, A. Viccellio, K. Mueller, "A Visual Analytics Framework for Emergency Room Clinical Encounters," *IEEE VisWeek Workshop on Visual Analytics in Health Care*, 2010.

[13] Google Body Browser: http://bodybrowser.googlelabs.com

Engaging Clinicians in the Visualization Design Process – Is It Possible?

Kostas Pantazos

IT-University of Copenhagen

Abstract

Creating and customizing visualization for electronic health record data requires a close collaboration with clinicians, to understand their tasks, needs and mental model. This process can develop into an infinite process. Taking into consideration the existence of clinicians with advanced IT knowledge, but not programmers, we focus on engaging them to create their own visualizations. This paper presents how clinicians can use uVis Studio to create three visualizations by dragging and dropping controls into the design panel, and specifying formulas for each control in the property grid.

Keywords: Visualization Tool, Spreadsheet Formulas, Development Environment, Design Process, Health Care.

Index Terms: H.5.2. [Information Interfaces & Presentation]: User Interfaces – Graphical User Interfaces (GUI)

1 Introduction

Healthcare systems provide a huge amount of data and the challenge of presenting these data is present. Clinicians need easy and intuitive presentations that fulfill their tasks and needs based on their experience and knowledge [7]. Most of EHR systems use more table or text based presentation rather than visualization techniques. Innovative visualizations like LifeLines [9], TimeLine [3], etc. provide a better presentation. These visualizations have been developed in close collaboration between developers and clinicians who have the domain knowledge. Creating and customizing advanced visualizations need programming skills and considerable time.

Although several visualizations have been developed for clinical data, there is a need for more novel and customizable visualizations [3]. Clinicians need presentations which are easy to understand and to access the right information [3]. Furthermore, the visualization has to match the mental model of the clinician. To overcome this challenge, it is recommended that clinicians are involved during the development process of a user interface or visualization [7]. Applying user-centered design may resolve these issues, but still questions like: "What about the clinicians that did not participate in the design process? Are the representatives a good sample, to conclude to the right visualization?". Furthermore, is the same visualization sufficient for the same department but in different hospitals ? Answering these questions raises several challenges which are also closely related with the available time, budget and resources used.

Using user-centered design does not solve the problem of customizability; adjusting an existing visualization to clinician needs. For instance, different departments or different hospitals have different needs. Different clinicians perform the same tasks in different ways, because of different experiences, knowledge and so forth. The same visualizations can be integrated in different departments or used by different clinicians, but to achieve better user satisfaction some changes may be needed. Furthermore, there is a need for more customizable visualizations to fulfill users' needs [3], and more tools which can support this customizability.

Nowadays, some clinicians have gained advanced IT skills, starting from simple browsing through web-applications to more advanced applications, such as MS Excel or MS Access. For instance at Bispebjerg hospital in Copenhagen, Denmark, a department uses a system developed in MS Access by one of the clinicians. We believe that in the healthcare environment there are a considerable number of such clinicians with advanced IT knowledge. So, with proper training, engaging clinicians in the process of developing their own visualizations using a specialized development environment will increase even more the possibility of developing successful visualizations for clinical data.

We present uVis, a formula-based visualization tool for clinicians. This tool provides clinicians with a development environment (uVis Studio) to design their visualizations. Clinicians with advanced spreadsheet level knowledge and familiar with basic database concepts can design visualization by dragging and dropping controls into the design panel. Next, specifying simple and advanced formulas in the property grid, they can bind controls to data and specify controls properties such as color, height, width, etc. The uVis Studio provides the basic features a development environment has, and more specialized ones such as data related intellisense and a design panel which shows visualizations as it would look to the end-users, described in the next sections. Finally, clinicians without IT experience can collaborate with IT experienced clinicians to create visualizations and use them as well.

2 Related work

2.1 Visualizations in healthcare

One of the most well-known visualizations in healthcare is the LifeLines [9]. It presents the history of a patient's medical record and it was designed in close collaboration with clinicians initially, and later with a cardiologist. This presentation uses the timeline metaphor, data presented in facets, color coding and size coding. The evaluation showed that the Lifelines was more understandable and that clinicians responded faster than the traditional presentations. This visualization was developed in Java, and customizing it requires advanced programming skills. The TimeLine system by Bui et. al. visualizes problem-centric patient data [3]. Their study showed that clinicians need more flexible visualizations which fulfill their needs and tasks. A need for more flexible visualization and customizable by clinicians is raised by An et. al. [1]. An integrated viewer for EHR was developed with basic visualization techniques, where clinicians were able to hide and show visualizations but not customize them to their needs.

Although, several previous research projects have concluded that there is a need for more customizable visualizations in healthcare, to our knowledge there is no previous research addressing this problem or engaging clinicians directly in the development process.

2.2 Visualization tools

We investigated some popular tools in the market for non-programmers mainly used in the business area. MS Excel [8] provides a user-friendly interface where built-in visualizations can be created with few steps. However, this tool provides a limited number of visualizations which are not fully customizable. For instance, graph colors cannot reflect data values. Furthermore, users cannot create new visualization types and integrate them into MS Excel. Finally, due to the amount and structure of data in an EHR system, clinicians may encounter difficulties in creating meaningful visualizations with MS Excel. Other visualization tools such as Spotfire [10] and Tableau [11] are more specialized in data visualization and provide a larger variety of visualizations. Nevertheless, these tools do not support users to create and customize advanced visualizations, such as LifeLines. User creativity is restricted to the pre-designed views. Furthermore, creating appropriate visualizations with Tableau or Spotfire needs some advanced knowledge on how to create visualizations.

In academia, there are several visualization toolkits [2, 5, 6] for programmers. Programmers can create and customize visualizations by means of programming. Unfortunately, this approach is too complex for users with advanced spreadsheet-like knowledge, such as clinicians. Most of these toolkits miss an integrated development environment. Usually, they can be integrated in general-purpose integrated development environments (IDE) such as Visual Studio, Eclipse, etc., but still is not enough for non-programmers. A specialized IDE should support users in creating and customizing visualizations by means of simple actions such as drag-and-drop.

3 SOLUTION

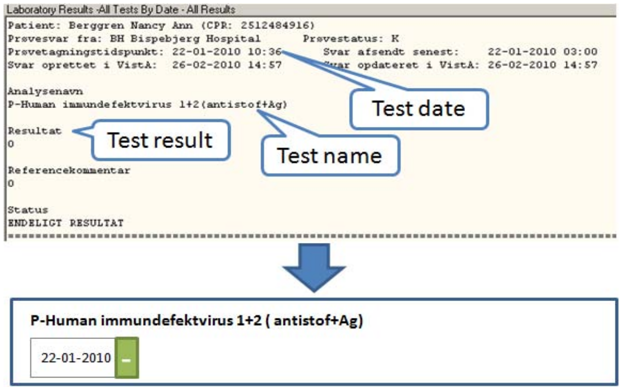
Previous research [1, 3] has been using user-centric design where clinicians had a close collaboration with the developer. We propose a different approach on developing visualizations for healthcare data: allow clinicians with advanced IT knowledge to create and customize their own visualizations using uVis.

uVis Studio (figure 2) is the development environment of uVis and contains six work areas. *Toolbox* lists the available controls, and supports drag-and-drop. *Design Panel* shows the visualization as it would look to the end-user. This panel is updated every time a control is dragged-dropped or a control property is changed. Hence, the user sees exactly the same screen in development mode as well as in end-user mode. *Property-Grid* is the area where a user can type the formulas. We integrated the intellisense feature in the Property-Grid to reduce typing errors and misunderstandings. Furthermore, the intellisense assists clinicians with suggestion related to control properties, tables and table fields. *Solution Explorer* is the area where project files are listed. The clinician can create a new project by adding a visualization mapping document (.vism) and a visualization file (.vis). Vism files contain information regarding the database the user is using, the tables, etc. The Vis file contains the visualization specifications. *Design Modes* allows the user to choose the mode for viewing and interacting with visualizations in the design panel. For instance, the user can select the mode *InteractionMode*, which deactivates event handlers attached to the visualization in the development environment. *Data Map*, currently under development, provides a visual overview of tables, fields and relationships in the database the user is using. It resembles an entity relationship (ER) diagram.

In the remainder of this section, we present three scenarios, three visualizations and elaborate on how they were created by the author.

3.1 Scenario 1: Simple LabResults visualization

In one of the clinics at Copenhagen Hospital, clinicians use the VistA EHR system. For each patient that comes in the clinic, they have to check the lab results of the patient. Figure 1.a presents a screenshot of the presentation of a patient lab result that clinicians use, and our simple solution using uVis in figure 1.b. Clinicians have to go through all the cumbersome texts for more than one lab test and find the important information for the patient. The lab test has a positive or negative result. A simple overview of the current state of the patient is missing. In the early phase of our research, we collaborated with clinicians who identified three important variables (date, result and lab name) in the texts, which are used in our visualization created using uVis Studio. Our approach is trying to minimize this collaboration and empower the clinicians to create their visualization.


Laboratory Results -All Tests By Date - All Results

```
Patient: Berggren Nancy Ann (CPR: 2512484916)
Prøvesvar fra: BH Bispebjerg Hospital       Prøvestatus: K
Prøvetagningstidspunkt: 22-01-2010 10:36      Svar afsendt senest:     22-01-2010 03:00
Svar oprettet i VistA:  26-02-2010 14:57      Svar opdateret i VistA: 26-02-2010 14:57

Analysenavn
P-Human immundefektvirus 1+2(antistof+Ag)

Resultat
0

Referencekommentar
0

Status
ENDELIGT RESULTAT
===============================================================================
```

P-Human immundefektvirus 1+2 (antistof+Ag)

22-01-2010

Figure 1. a) Current presentation at the clinic and b) a potential solution for presenting patient Lab Results.

Preconditions: uVis can visualize only relational data at the moment, for instance data in MS Access. The Vism file has to be created the first time by the database manager, unless the clinician who will use uVis studio has good database knowledge. Furthermore, an introduction of how the studio works and how to use formulas is necessary for clinicians.

3.1.1 Using uVis Studio

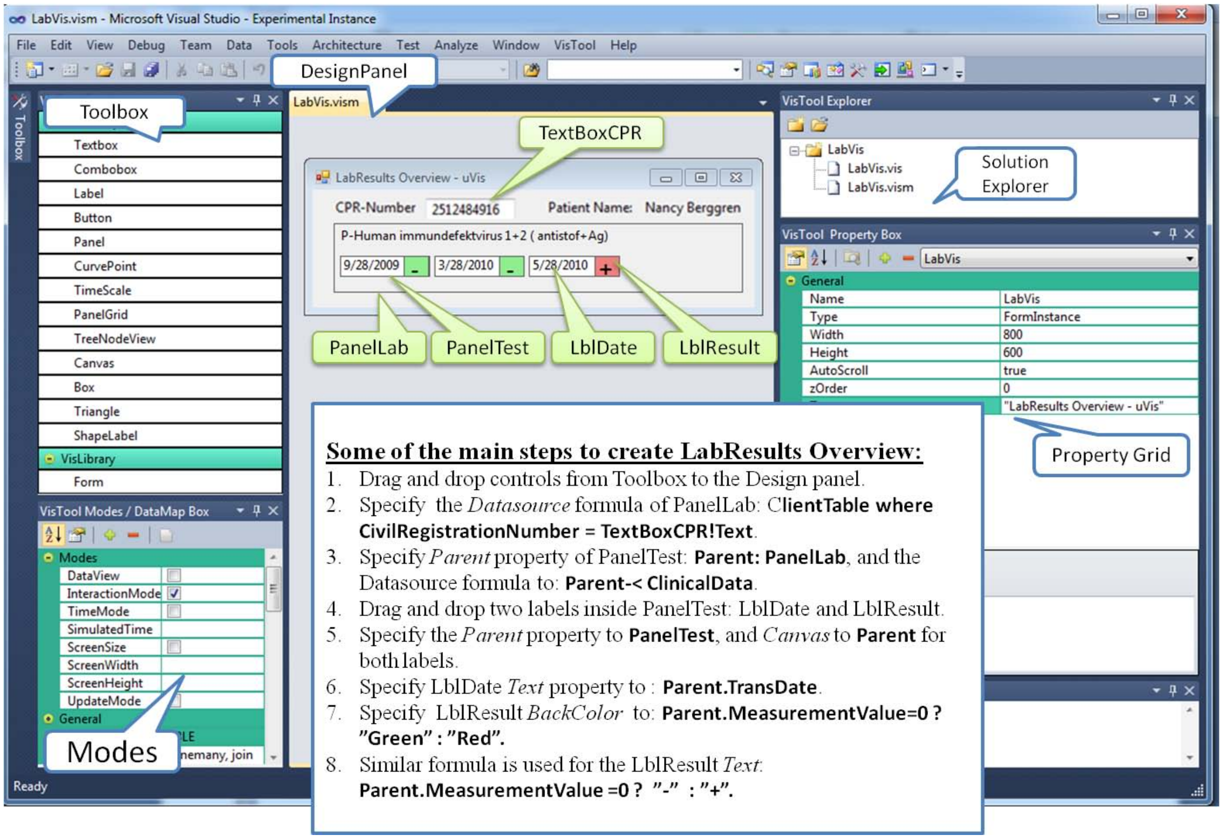
Figure 2 shows a screenshot of the studio, containing a simple visualization for the lab results, and some of the steps clinicians have to follow. The clinician opens the uVis Studio and selects the Vism file using the explorer. The default Vis file is opened in the design panel. In our case it will be an empty form.

Clinicians can drag and drop controls (e.g. panel, label, textbox, etc.) in the design panel. Furthermore, they can resize the controls and move them around the design panel. For each control they specify simple and advanced formulas for control properties in the property grid. Every change done in the property grid reflects on real-time on the design panel. Unlike other development environment, uVis Studio shows the form exactly as it will be shown at the end-user outside Studio. Clinicians use the property grid to specify the formulas. Intellisense feature helps them to write the correct formulas. For instance, clinician starts typing "cli" in the DataSource property and a list of suggestions will pop-up with name of tables, table fields, controls and control properties that contain "cli".

3.1.2 Key Principles of uVis Kernel

In this section we present some of the key principles of uVis Kernel which are used in creating the LabResult visualization, Figure 2.

Controls: Visualizations are created by combining .Net controls, simple shapes (e.g. triangle) and several special uVis controls (e.g. timescale). A control can be bound to data that



CPR-Number 2512484916 Patient Name: Nancy Berggren

LabVis

Name LabVis
Type FormInstance
Width 800
Height 600
AutoScroll true
zOrder 0
"LabResults Overview - uVis"

Some of the main steps to create LabResults Overview:

1. Drag and drop controls from Toolbox to the Design panel.
2. Specify the *Datasource* formula of PanelLab: C**lientTable where CivilRegistrationNumber = TextBoxCPR!Text**.
3. Specify *Parent* property of PanelTest: **Parent: PanelLab**, and the Datasource formula to: **Parent-< ClinicalData**.
4. Drag and drop two labels inside PanelTest: LblDate and LblResult.
5. Specify the *Parent* property to **PanelTest**, and *Canvas* to **Parent** for both labels.
6. Specify LblDate *Text* property to : **Parent.TransDate**.
7. Specify LblResult *BackColor* to: **Parent.MeasurementValue=0 ? ”Green” : ”Red”.**
8. Similar formula is used for the LblResult *Text*: **Parent.MeasurementValue =0 ? ”-” : ”+”.**

Figure 2. Creating LabResults overview with uVis Studio.

makes it repeat itself. A control has a number of properties that specify its appearance and its behavior.

Formulas: Control properties can be specified by spreadsheet-like formulas. The formula specifies how to compute a property value for a control. A formula can refer to data in the database, control properties. uVis kernel computes the formulas for each control, and sets the property values accordingly.

Bind control to data: Each control may have a data source that binds it to data rows. To define the data source, in this case the clinician specifies the DataSource, the uVis property of the control. The clinician writes a formula which represents an SQL statement. uVis kernel translates the DataSource formula into an SQL statement, retrieves data from the database and generates the corresponding record set. Next, the control creates one control for each row in the record set. Each control is bound to a row in the record set.

To create the visualization showed in figure 2, we used only two tables from our EHR database: ClientTable and ClinicalData. Each patient may have one or more clinical data. For instance in Figure 2, the patient is tested three times for P-Human immundefektvirus 1+2.

The clinician specifies the DataSource of panel PanelLab as follows:

ClientTable where CivilRegistrationNumber = TextBoxCPR!Text

ClientTable refers to a table in the data model and CivilRegistrationNumber is a field in table ClientTable. The dot (.) operator allows the clinician to access a table field. TextBoxCPR is the control of type TextBox that shows the patient civil registration number (CPR). The operator ! allows the user to access a control property. Thus, TextBoxCPR!Text is the current patient’s CPR. As a result, the data source of PanelLab is the patient record whose civil registration number is specified in TextBoxCPR. As a result, uVis kernel creates one PanelLab control.

To show the lab tests of a patient, the user drags and drops a panel (PanelTest) inside PanelLab and specifies the DataSource of PanelTest as follows:

Parent -< ClinicalData.

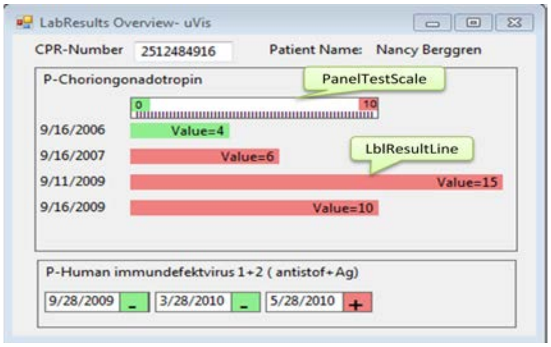
Parent means the data parent of PanelTest, in this case PanelLab. The operator -< allows us to navigate from one row to multiple rows. Therefore, we navigate from the parent row (the ClientTable row) to the related ClinicalData rows. This allows us to access the lab tests of the patient. uVis kernel automatically detects the tables and table fields used in the formulas. Next, uVis Kernel translates the formula to an SQL statement, which is executed and a record set is created. In this case the record set contains three rows. Clinicians are not involved in this process, apart from the fact that they need to specify the correct formula in the property grid.

3.2 Scenario 2: Advanced LabResults visualization

We present in addition lab tests with numerical value as results. Instead of going through the text, clinicians can create or customize the first version of Lab Results Overview and present numerical lab tests as shown in Figure 3.

Following the same principles presented before, the clinician can bind controls to data. PanelTestScale presents visually the lowest and highest value this test may have in theory. However, in this case one of the test result was higher than 10. In this presentation, the clinician can spot it out easily, compared to the text based presentation. To align LblResultLine to PanelTestScale clinician specifies the Left property to:

PanelTestScale!Left



LabResults Overview- uVis

CPR-Number 2512484916 Patient Name: Nancy Berggren

P-Choriongonadotropin

9/16/2006	Value=4
9/16/2007	Value=6
9/11/2009	Value=15
9/16/2009	Value=10

P-Human immundefektvirus 1+2 (antistof+Ag)

9/28/2009 - 3/28/2010 - 5/28/2010 +

Figure 3. LabResults Overview using uVis Studio

To calculate the width of the LblResultLine, the clinician specifies this formula for the width property:

PanelTestScale!Width * Me.MeasurementValue / (Me.ValueHigh - Me.ValueLow).

Me is used to refer to the current instance, which is bound to a row. Using the dot operator we can navigate to a specific field of this row (MeasuremnetValue, ValueHigh and ValueLow in our case).

3.3 Scenario 3: LabResults using LifeLines

In the last scenario, the clinician creates a simple LifeLines visualization for some of the lab tests, shown in figure 4.

The clinician follows the same steps as before to bind controls to data. The difference in this case is the Timescale control, which is a uVis control. The clinician defines the period shown in the timescale by specifying the BorderValues to:

#2011-08-01#, #2011-09-01#

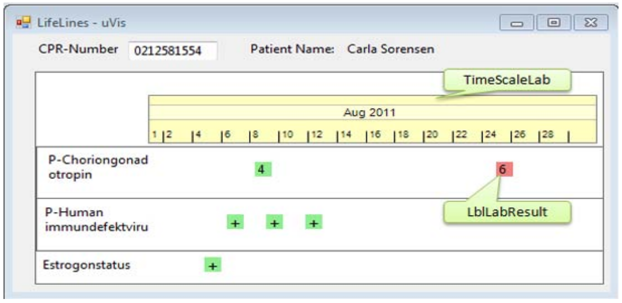
Clinicians can interact with the TimeScale control, moving the date backwards or forwards. To align the LblLabResult the clinician specifies the left position to:

TimeScaleLab!HPos (me.TransDate).

HPos is a special function in the timescale which translate date to pixels.

4 Discussion

Nowadays, computers are part of our daily and working life. More and more users are using computers to facilitate their working process. Starting from simple usage (such as checking emails, browsing web application), users, especially the new generation, are moving towards a better and broader understanding of how to utilize computers in daily work. The real case in Copenhagen Hospital, where a clinician developed an application in MS Access, confirms this tendency. Although several visualization tools exist, there is a need for new tools which provide a development environment for clinicians with advanced IT knowledge, but not programmers. Such a tool will facilitate the development process, allowing clinicians to create and customize their own visualization based on the department needs or their mental model.



LifeLines - uVis

CPR-Number 0212581554 Patient Name: Carla Sorensen

Figure 4. Simple LifeLines visualization using uVis Studio

In this paper, we present an on-going research project, which focuses on engaging clinicians in developing simple and advanced visualization using spreadsheet-like formulas. The spreadsheet formulas have proven to be successful approach among users and programmers [4]. Furthermore, by means of the development environment, clinicians can customize their visualization and adjust them to fulfill their needs.

The abovementioned visualizations were created by the author who has a good understanding of uVis Studio and formula principles, but is not a clinician. A more in depth evaluation with real clinicians is needed, and we are planning to conduct it in the future. The evaluation will show if our approach is adequate and if it is possible to engage clinicians in the visualization design process.

Now, we are focusing on making uVis Studio more stable. Data Map is being developed and simpler and advanced controls are being developed. A more specialized error messaging system for clinicians is being developed.

5 Conclusion

In this paper we presented a new visualization tool for clinicians. Clinicians can create and customize visualizations by means of iteratively dragging and dropping controls and specifying spreadsheet-like formulas. Although, three visualizations for lab results were developed, we plan to conduct an evaluation with real clinicians. To conclude, in this paper we present a first attempt to engage clinicians more and allow them to visualize the data in their own way.

References

[1] J. An, Z. Wu, H. Chen, X. Lu, H. Duan: Level of detail navigation and visualization of electronic health records, Proceedings of Biomedical Engineering and Informatics (BMEI), 2010.

[2] M. Bostock and J. Heer. Protovis: A graphical toolkit for visualization. IEEE Trans. Vis. and Comp. Graphics, 15(6):1121–1128, 2009

[3] A.A. Bui, D.R. Aberle, and H. Kangarloo, "TimeLine: Visualizing Integrated Patient Records", IEEE transactions on information technology in medicine, vol. 11, no. 4, 2007.

[4] M. Burnett, John Atwood, Rebecca Walpole Djang, James Reichwein, Herkimer Gottfried, and Sherry Yang. 2001. Forms/3: A first-order visual language to explore the boundaries of the spreadsheet paradigm. J. Funct. Program. 11, 2 , 155-206, March 2001.

[5] Flare. http://flare.prefuse.org, February 2011.

[6] J. Heer, S. K. Card, and J. A. Landay. "prefuse: a toolkit for interactive information visualization". *In Proc. ACM CHI*, pages 421–430, 2005

[7] C. M. Johnson, T. R. Johnson, J. Zhang. 2005. A user-centered framework for redesigning health care interfaces. J. of Biomedical Informatics 38, 1, 75-87, February 2005.

[8] Microsoft Excel. http://office.microsoft.com/en-us/excel/, February 2011.

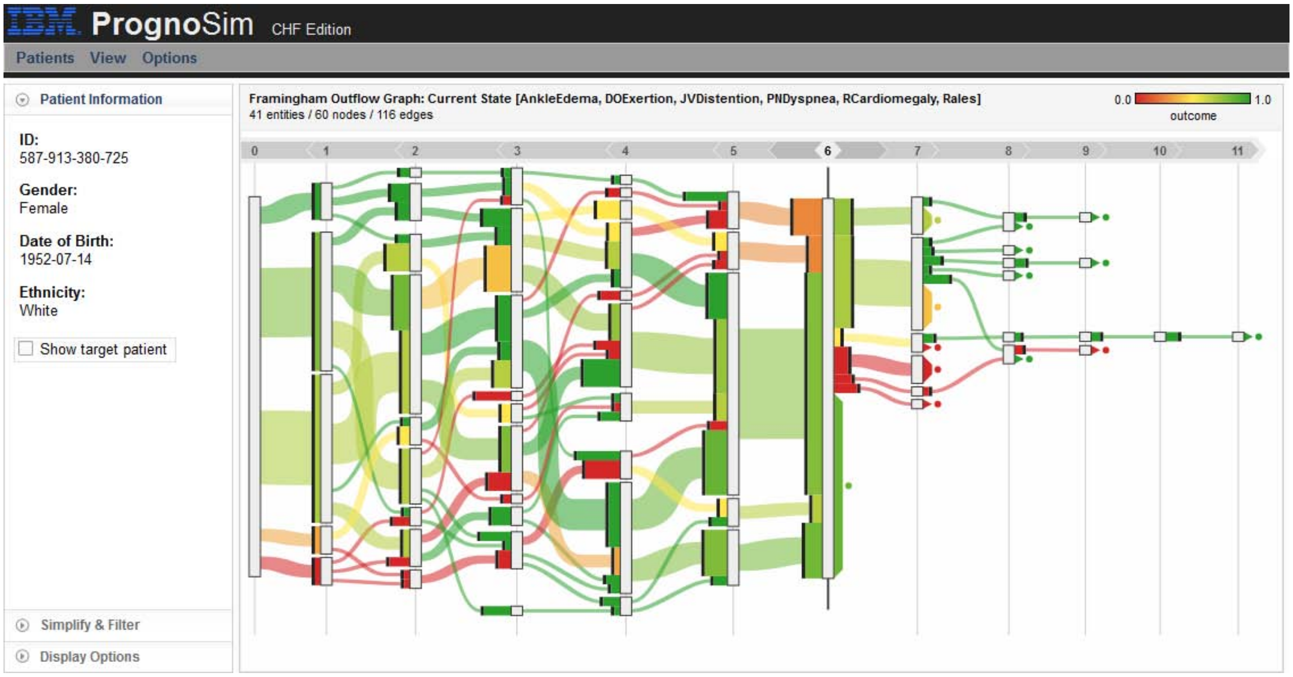
[9] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman, "LifeLines: Using visualization to enhance navigation and analysis of patient records," *In Proc. American Medical Informatic Association Annu. Fall Symp*., Orlando, FL, , pp. 76-80, November 1998.

[10] Spotfire. http://spotfire.tibco.com/, February 2011.

[11] Tableau. http://www.tableausoftware.com/, February 2011.

Outflow: Visualizing Patient Flow by Symptoms and Outcome

Krist Wongsuphasawat, and David H. Gotz



ID:
587-913-380-725

Gender:
Female

Date of Birth:
1952-07-14

Ethnicity:
White

□ Show target patient

Framingham Outflow Graph: Current State [AnkleEdema, DOExertion, JVDistention, PNDyspnea, RCardiomegaly, Rales]
41 entities / 60 nodes / 116 edges

Fig. 1. Outflow aggregates temporal event data from a cohort of patients and visualizes alternative clinical pathways using color-coded edges that map to patient outcome. Interactive capabilities allow users to explore the data and uncover insights.

Abstract—Electronic Medical Record (EMR) databases contain a large amount of temporal events such as diagnosis dates for various symptoms. Analyzing disease progression pathways in terms of these observed events can provide important insights into how diseases evolve over time. Moreover, connecting these pathways to the eventual outcomes of the corresponding patients can help clinicians understand how certain progression paths may lead to better or worse outcomes. In this paper, we describe the Outflow visualization technique, designed to summarize temporal event data that has been extracted from the EMRs of a cohort of patients. We include sample analyses to show examples of the insights that can be learned from this visualization.

Index Terms—Outflow, Information Visualization, Temporal Event Sequences, State Diagram, State Transition

1 Introduction

Electronic medical records (EMRs) are proliferating throughout the healthcare system. At major medical institutions such as hospitals and large medical groups, these computer-based systems contain vast amounts of historical patient data complete with patient profile information, structured observational data such as diagnosis codes and medications, as well as unstructured physician notes. The information in these enormous databases can be useful in guiding the diagnosis of incoming patients or in clinical studies of a disease. However, the vast amount of information can be overwhelming and makes these datasets difficult to analyze. In particular, EMR databases contain a large amount of temporal disease events such as diagnosis dates and the onset dates for various symptoms. Analyzing disease progression pathways in terms of these observed events can provide important insights into how diseases evolve over time. Moreover, connecting these pathways to the eventual outcomes of the corresponding patients can help clinicians understand how certain progression paths may lead to better or worse outcomes.

In this paper, we describe the Outflow visualization technique. Outflow is designed to summarize temporal event data that has been extracted from the EMRs of a cohort of patients. We present a novel interactive visual design which combines multiple patient records into a graph-based visual presentation. Users can manipulate the visualization through direct interaction techniques (e.g., selection and brushing) and a series of control widgets. The interactions allow users to explore the data in search of insights. Throughout the paper we describe Outflow using a motivating problem related to the diagnosis of congestive heart failure. We include two sample analyses to show examples of the insights that can be learned from this visualization.

The rest of the paper are organized as follows. We describe our motivating problem in Section 2 and review related work in Section 3. We explain the design of Outflow in Section 4 and demonstrate preliminary analyses in Section 5. The paper concludes in Section 6.

- *Krist Wongsuphasawat is with University of Maryland. This work is part of his internship at IBM T.J. Watson Research Center. , E-mail: kristw@cs.umd.edu.*
- *David H. Gotz is with IBM T.J. Watson Research Center, E-mail: dgotz@us.ibm.com.*

Manuscript received 31 March 2011; accepted 1 August 2011; posted online 23 October 2011; mailed on 14 October 2011.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org.

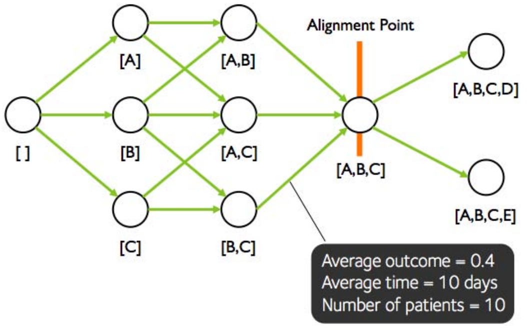


Fig. 2. Multiple medical records are aggregated into a representation called an Outflow graph. This structure is a directed acyclic graph (DAG) that captures the various event sequences that led to the alignment point and all the sequences that occurred after the alignment point. Aggregate patient statistics are then anchored to the graph to describe specific patient subpopulations.

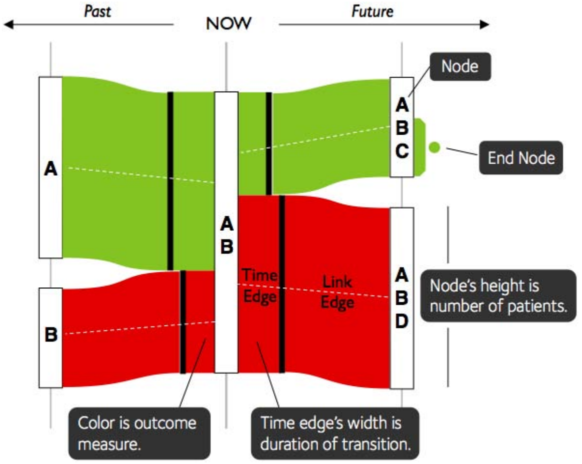

Past
NOW
Future
Node
End Node
A
B
A B
A B C
A B D
Time Edge
Link Edge
Node's height is number of patients.
Color is outcome measure.
Time edge's width is duration of transition.

Fig. 3. Outflow visually encodes nodes in the Outflow graph using rectangles while edges are represented using two distinct visual marks: time edges and link edges. Color is used to encode average outcome.

2 Motivating Problem

Congestive heart failure (CHF) is generally defined as the inability of the heart to supply sufficient blood flow to meet the needs of the body. CHF is a common, costly, and potentially deadly condition that afflicts roughly 2% of adults in developed countries with rates growing to 6-10% for those over 65 years of age [12]. The disease is difficult to manage and no system of diagnostic criteria has been universally accepted as the gold standard.

One commonly used system comes from the *Framingham study* [11]. This system requires the simultaneous presence of at least two major symptoms (e.g., S3 gallop, Acute pulmonary edema, Cardiomegaly) or one major symptom in conjunction with two minor symptoms (e.g., Nocturnal cough, Pleural effusion, Hepatomegaly). In total, 18 distinct Framingham symptoms have been defined.

While these symptoms are used regularly to diagnose CHF, our medical collaborators are interested in understanding how the various symptoms and their order of onset correlate with patient outcome. To examine this problem, we were given access to an anonymized dataset of 6,328 patient records. Each patient record includes timestamped entries for each time a patient was diagnosed with a Framingham symptom. For example:

Patient#1:(27 Jul 2009, Ankle edema), (14 Aug 2009, Pleural effusion), ...

Patient#2:(17 May 2002, S3 gallop), (1 Feb 2003, Cardiomegaly), ...

In line with the use of Framingham symptoms for diagnosis, we assume that once a symptom has been observed it applies perpetually. We therefore filter the event sequences for each patient to select only the first occurrence of a given symptom type. The filtered event sequences describe the *flow* for each patient through different disease states. For example, a filtered event sequence ***symptom A*** → ***symptom B*** indicates that the patient's flow is ***no symptom*** → ***symptom A*** → ***symptoms A and B***. The data also has an outcome for each patient (dead (0) or alive (1)).

Our analysis task, therefore, is to examine aggregated statistics for the flows of many patients to find common disease states and transitions between states. In addition, we wish to discover any correlations between these paths and patient outcome.

3 Related Work

3.1 Temporal Event Sequence Visualizations

Many researchers have explored visualization techniques for temporal event sequences. In the early years, many systems focused on visualizing a single record [1, 2, 6, 8, 9, 16]. The most common approach is to place the events on a horizontal timeline according to the time that events occurred. Later, attention shifted towards visualizing multiple records in parallel. One popular technique is to stack instances of single-record visualizations and to provide additional functionality for searching [7, 21, 22, 23, 26], filtering [23], and grouping [5, 14]. However, these approaches do not aggregate nor provide any abstraction of multiple event sequences. Most recently, a technique called *LifeFlow* [25] introduced a way to aggregate and provide an abstraction for multiple event sequences. However, LifeFlow's aggregation combines multiple event sequences into a tree, while Outflow's aggregation combines multiple event sequences into a graph.

3.2 State Diagram Visualizations

Our approach aggregates event sequences into an *Outflow graph* which is analogous to a state diagram [4] or state transition graph. State diagrams are used in computer science and related fields to represent a system of states and state changes. State diagrams are generally displayed as simple node-link diagrams where each state is depicted as a node and transitions are drawn as links [3]. Many visualizations of state diagrams have been developed [3, 17, 18, 20, 24]. These typically focus on multivariate graphs where a number of attributes are associated with every node. Some support exploration of sequences of three or more states. Variants on traditional state diagrams have also been explored, such as *Petri nets* (also known as a *place/transition* net or *P/T net*) [13] which offer a graphical notation for stepwise processes that include choice, iteration, and concurrent execution. However, to the best of our knowledge, these approaches do not display or allow easy comparison of the transition time, which is one of Outflow's design goals.

3.3 Flow & Parallel Coordinates Visualizations

Another group of visualizations called Sankey Diagrams [19] was designed to visualize flow quantities in process systems. However, they only focus on displaying the proportion of the flow that splits in different ways, without temporal information. The visual display of Outflow also looks similar to parallel coordinates [10], but the underlying data types are different. Parallel coordinates are used for categorical data while Outflow was designed for temporal event sequences.

4 Description of the Visualization

4.1 Data Aggregation

The first step in Outflow is data aggregation. We begin by selecting an *alignment point*. For example, we can align a set of patient event sequences around a state where all patients have the same three symptoms A, B and C and no other symptoms. After choosing an alignment point, we construct an *Outflow graph* (Figure 2) using data from all patients that satisfy the alignment point.

The Outflow graph is a state diagram represented using a directed acyclic graph (DAG). The states are the unique combinations of symptoms that were observed in the data. Edges capture symptom transitions. Each edge is annotated with the number of patients that make the corresponding transition, the average time gap between the states, and the average outcome of the patient group.

Therefore, the Outflow graph captures all event paths that led to the alignment point and all event paths that occur after the alignment point. Our prototype implementation lets users select a target patient from the database and uses the target patient's current state as the alignment point. This approach allows for the analysis of historical data when considering the possible future progression of symptoms for the selected target patient.

4.2 Visual Encoding

Based on the information contained in the Outflow graph, we have designed a rich visual encoding that displays (a) the time gap for each state change, (b) the cardinality of patients in each state and state transition, and (c) the average patient outcome for each state and transition. Drawing on prior work from FlowMap [15] and LifeFlow [25], we developed the visual encoding shown in Figure 3.

Node (State): Each node is represented by a rectangle which has its height proportional to the number of patients.

Layer: We slice the graph vertically into layers. Layer *i* contains all Outflow graph nodes with *i* symptoms. The layers are sorted from left to right, showing information from the past to the future. For example, in Figure 1, the first layer (layer 0) contains only one node, which represents patients that have no symptom. The next layer (layer 1) has five nodes, one for each first-occurring symptom in the patient cohort.

Edge (Transition): Each edge is displayed using two visual marks: a *time edge* and a *link edge*. Time edges are rectangles that whose width is proportional to the average time gap of the transition and height is proportional to the number of patients. Link edges connect nodes and time edges to convey sequentiality.

End Node: Each patient's path can stop in a different state. We use a trapezoid followed by a circle to mark these points. The height of the trapezoid is proportional to the number of patients whose path stops at a given point.

Color-coding: Colors assigned to edges and end nodes are used to encode the average outcome for the corresponding set of patients. The color scales linearly from red to green with red representing the worst and green representing the best outcomes.

4.3 Interactions

To allow interactive data exploration, we further designed Outflow to support the following user interaction capabilities.

Panning & Zooming: Users can pan and zoom to uncover detailed structure.

Filtering: Users can filter both nodes and edges based on the the number of associated patients to remove small subgroups.

Symptom Selection: Users can select which symptom types are used to construct the Outflow graph. This allows, for instance, for the omission of symptoms that users deem uninteresting. For example, a user can remove *Nocturnal Cough* if they deem it irrelevant to an analysis and the visualization will be recomputed dynamically.

Brushing: Hovering the mouse over a node or an edge will highlight all paths traveled by patients passing through the corresponding point in the outflow graph (see Figure 4).

Tooltips: Hovering also triggers the display of tooltips which provide more information about individual nodes and edges. Tooltips shows all symptoms associated with the corresponding node/edge, the average outcome, and the total number of patients in the subgroup (see Figure 4).

5 Preliminary Analysis

We have integrated the Outflow visualization technique into a prototype decision support system for CHF patients called *PrognoSim*. This system uses a patient similarity-based approach to provide medical intelligence. PrognoSim is a web-based application written using Java's J2EE platform and Apache Tomcat as the application server environment. The PrognoSim user interface is rendered using HTML and JavaScript. Dojo is used for traditional user interface widgets. The Outflow visualization component is rendered on an HTML 5 canvas via a scenegraph-based JavaScript visualization library named CVL.

We used Outflow within PrognoSim to view the evolution over time for a cohort of CHF patients similar to a clinician's current patient. Our initial analysis illuminates a number of interesting findings and highlights that various types of patients evolve differently. We share two such evolution patterns as examples of the type of analysis that can be performed using the Outflow technique.

Leading Indicators. In several scenarios, patient outcome is strongly correlated with certain leading indicators. For example, consider the patient cohort visualized in Figure 1. The strong red and green colors assigned to the first layer of edges in the visualization shows that the eventual outcome for patients in this cohort is strongly correlated with the very first symptom to appear. Similarly, the strong red and green colors assigned to the first layer of edges after the alignment point show that the next symptom to appear may be critical in determining patient outcome.

Progressive Complications. In contrast to the prior example, which showed strong outcome correlation with specific paths, the patient cohort in Figure 5 exhibits very different characteristics. At each time step, the outcomes across the different edges are relatively equal. However, the outcomes transition from green to red when moving left to right across the visualization. This implies that for this group of patients, no individual path is especially problematic historically. Instead, a general increase in co-occurring symptoms over time is the primary risk factor.

6 Conclusions and Future work

We have introduced a novel visualization called Outflow that summarizes temporal event data extracted from multiple patient medical records to show aggregate disease evolution statistics for a cohort of patients. We described our motivating problem in the study of congestive heart failure and presented the main visual design concepts behind our visualization. We also described a number of interactive features in Outflow that allow more sophisticated analyses. Finally, we briefly shared two example analysis results which highlight some of the capabilities of our approach.

Due to these early promising results, we plan to continue work on this topic in the future. We believe that there are many promising directions to explore including integration with forecasting/prediction algorithms, the use of more sophisticated similarity measures, and deeper evaluation studies with practitioners. Moreover, the flexibility of Outflow's design means it can be used beyond our motivating problem and can be useful for a range of medical (and non-medical) problems which involve temporal event data.

Acknowledgments

The authors wish to thank Charalambos (Harry) Stavropoulos, Robert Sorrentino and Jimeng Sun for their help and suggestions.

References

[1] W. Aigner, S. Miksch, B. Thurnher, and S. Biffl. PlanningLines: Novel Glyphs for Representing Temporal Uncertainties and Their Evaluation. In *Proc. International Conf. Information Visualization (IV)*, pages 457–463, 2005.

[2] R. Bade, S. Schlechtweg, and S. Miksch. Connecting time-oriented data and information to a coherent interactive visualization. In *Proc. Annual SIGCHI Conf. Human Factors in Computing Systems (CHI)*, pages 105–112, 2004.

[3] J. Blaas, C. P. Botha, E. Grundy, M. W. Jones, R. S. Laramee, and F. H. Post. Smooth graphs for visual exploration of higher-order state transitions. *IEEE Trans. Visualization and Computer Graphics*, 15(6):969–76, 2009.

[4] T. Booth. *Sequential Machines and Automata Theory.* , 1967.

[5] M. Burch, F. Beck, and S. Diehl. Timeline trees: visualizing sequences of transactions in information hierarchies. In *Proc. Working Conf. Advanced Visual Interfaces (AVI)*, pages 75–82, 2008.

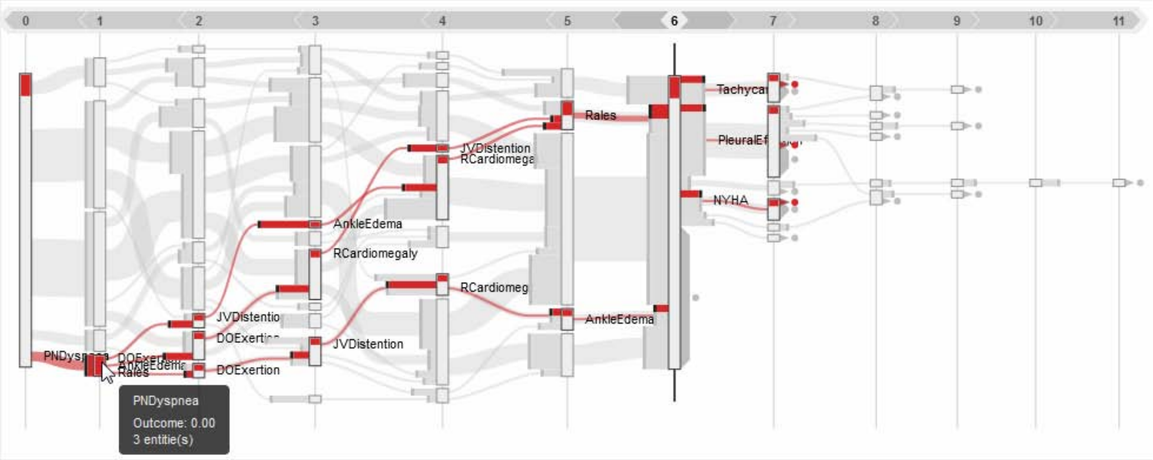


Fig. 4. Interactive brushing allows users to highlight paths emanating from specific nodes or edges in the visualization.

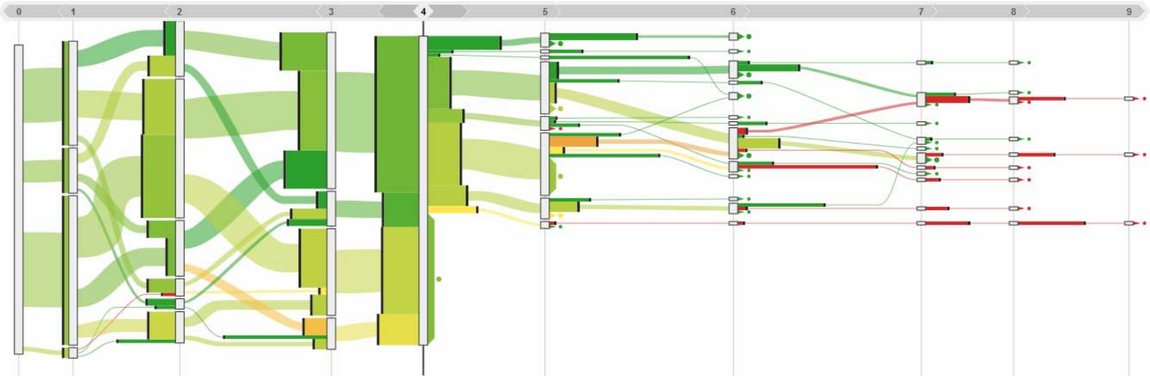


Fig. 5. The progression from green to red when moving left to right in this figure shows that patients with more symptoms exhibit worse outcomes.

[6] S. Cousins and M. Kahn. The visual display of temporal information. *Artificial Intelligence in Medicine*, 3(6):341–357, 1991.

[7] J. Fails, A. Karlson, L. Shahamat, and B. Shneiderman. A Visual Interface for Multivariate Temporal Data: Finding Patterns of Events across Multiple Histories. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST)*, pages 167–174, 2006.

[8] B. Harrison, R. Owen, and R. Baecker. Timelines: an interactive system for the collection and visualization of temporal data. In *Proc. Graphics Interface (GI)*, pages 141, 1994.

[9] G. Karam. Visualization using timelines. In *Proc. ACM SIGSOFT International Symp. Software Testing and Analysis*, pages 125–137, 1994.

[10] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: interactive exploration and visual analysis of categorical data. *IEEE Trans. Visualization and Computer Graphics*, 12(4):558–68, 2006.

[11] P. A. McKee, W. P. Castelli, P. M. McNamara, and W. B. Kannel. The natural history of congestive heart failure: the Framingham study. *The New England journal of medicine*, 285(26):1441–6, 1971.

[12] J. J. V. McMurray and M. A. Pfeffer. Heart failure. *Lancet*, 365(9474):1877–1889, 2005.

[13] C. A. Petri. Communication with automata. Technical report, DTIC Research Report, 1966.

[14] D. Phan, A. Paepcke, and T. Winograd. Progressive multiples for communication-minded visualization. In *Proc. Graphics Interface (GI)*, page 225, 2007.

[15] D. Phan, L. Xiao, and R. Yeh. Flow map layout. In *Proc. IEEE Symp. Information Visualization*, pages 219–224, 2005.

[16] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman. LifeLines: using visualization to enhance navigation and analysis of patient records. In *Proc. AMIA Annual Symp.*, pages 76–80, 1998.

[17] A. J. Pretorius and J. J. van Wijk. Visual analysis of multivariate state transition graphs. *IEEE Trans. Visualization and Computer Graphics*, 12(5):685–92, 2006.

[18] A. J. Pretorius and J. J. van Wijk. Visual Inspection of Multivariate Graphs. *Computer Graphics Forum*, 27(3):967–974, 2008.

[19] P. Riehmann, M. Hanfler, and B. Froehlich. Interactive Sankey diagrams. In *Proc. IEEE Symp. Information Visualization*, pages 233–240, 2005.

[20] F. van Ham, H. van de Wetering, and J. van Wijk. Interactive visualization of state transition systems. *IEEE Trans. Visualization and Computer Graphics*, 8(4):319–329, 2002.

[21] K. Vrotsou, K. Ellegard, and M. Cooper. Everyday Life Discoveries: Mining and Visualizing Activity Patterns in Social Science Diary Data. In *Proc. International Conf. Information Visualization (IV)*, pages 130–138, 2007.

[22] K. Vrotsou, J. Johansson, and M. Cooper. ActiviTree: interactive visual exploration of sequences in event-based data using graph similarity. *IEEE Trans. Visualization and Computer Graphics*, 15(6):945–52, 2009.

[23] T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman. Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proc. Annual SIGCHI Conf. Human Factors in Computing Systems (CHI)*, pages 457–466, 2008.

[24] M. Wattenberg. Visual exploration of multivariate graphs. In *Proc. Annual SIGCHI Conf. Human Factors in Computing Systems (CHI)*, page 811, 2006.

[25] K. Wongsuphasawat, J. Guerra Gómez, C. Plaisant, T. Wang, M. Taieb-Maimon, and B. Shneiderman. LifeFlow: Visualizing an Overview of Event Sequences. In *Proc. Annual SIGCHI Conf. Human Factors in Computing Systems (CHI)*, pages 1747–1756, 2011.

[26] K. Wongsuphasawat and B. Shneiderman. Finding comparable temporal categorical records: A similarity measure with an interactive visualization. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST)*, pages 27–34, 2009.

Clinical Applications of Star Glyphs and Ideas about Crowdsourcing Data Visualization Software

Jim DeLeo and James Cimino

National Institutes of Health Clinical Center
Bethesda Maryland

ABSTRACT

We describe our recent work with star glyph data visualization methods applied to clinical data derived from National Institutes of Health (NIH) clinical research protocols and we suggest a crowdsourcing approach for developing data visualization and computational intelligent software to mine data and discover new knowledge using clinical research data available through the NIH Biomedical Translational Medicine Informatics System (BTRIS).

KEYWORDS: Challenges, collaborative software development, competitions, crowdsourcing, data mining, data visualization, knowledge data discovery, parallel coordinates, star glyphs, software standards, radial plots.

INDEX TERMS: D.2.1 [Software Engineering]: Requirements/ Specifications – Elicitation methods; D.2.2 [Software Engineering]: Design Tools and Techniques – Software Libraries; G.4 [Mathematics of Computing]: Mathematical Software – Algorithm design and analysis; H.5.2 [Information Interfaces and Presentation]: User Interfaces-Graphical user interfaces (GUI); I.3.4 [Computer Graphics]: Graphics Utilities - Software support

1 INTRODUCTION

Data visualization methods can help us see and understand relationships in large multifactorial data arrays. They can also assist us in detecting patterns and anomalies not obvious with other forms of data representation. Data visualization methods are becoming increasingly popular for data exploration, data mining, information retrieval, and hypotheses suggestion in many different subject matter domains.

Our interest in data visualization has grown from our work in applying data visualization methods (particularly star glyphs and interactive parallel coordinates) to NIH clinical research protocol data. We believe these methods have good potential for catalyzing new medical knowledge insights and for producing informative data patterns that suggest hypotheses worthy of exploring. We now want to develop production quality software with good graphical user interfaces and good interfaces to archived data sources in order to expand our data visualization work and to provide extended computational support for biomedical researchers. We are in an ideal position at the NIH to develop and showcase this kind of software and to put it into practical use in supporting many of the more than a thousand clinical research protocols active here.

When star glyphs and other data visualization methods are made available in standardized, well-documented, easy-to-use readily-available software they should become important tools for gaining new insights and knowledge in medicine and other disciplines.

Here we show some of our current work in applying star glyph data visualization methods to clinical research data derived from NIH clinical research protocols. We also suggest a crowdsourcing approach to develop data visualization and computational intelligent software to mine data and discover new knowledge by using the clinical research protocol data available through the NIH Biomedical Translational Medicine Informatics System (BTRIS).

2 STAR GLYPH BASICS

Glyphs represent data values as shapes, textures and color attributes of graphical symbols [1, 2]. Many glyph representations have been proposed over the years including star glyphs [3], Andrews glyphs, [4], Chernoff faces [5], stick figure icons [6], shape coding [7] and DeLeo's star glyph movies [8]. Star glyphs (also known as radial plots) represent data values in the form of a star. Figure 1 illustrates a basic coordinate system frame for constructing a star glyph. This coordinate system indicates that there will be 20 variables plotted and that each variable will be scaled to the 0-1 interval and plotted on one of the spokes (rays) in the star glyph frame. Note that 0 corresponds to the center of the figure and 1 corresponds to the end of a spoke. The variables can be comprised of any mix of continuous and categorical variables. Any reasonable number of variables can be plotted. The ordering of the valuables is arbitrary and may be selected

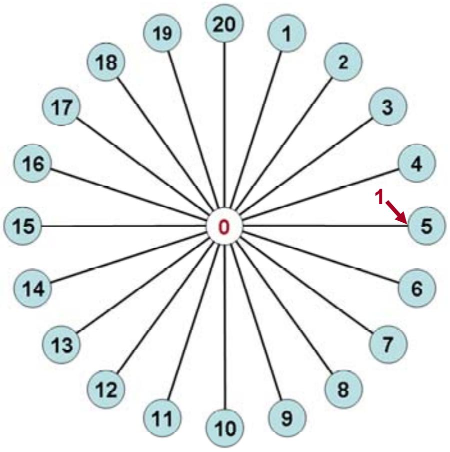


Figure 1. A coordinate system frame for constructing star glyphs.

National Institutes of Health Clinical Center
10 Center Drive
Bethesda, Maryland 20892
E-mail: jdeleo@nih.gov

according to attributes peculiar to the specific kind of data being plotted. There are many alternative ways to scale data onto the 0-1 as will be discussed below.

3 OUR EXPERIENCE WITH STAR GLYPHS

Here we present examples of the work we have been doing with star glyphs in clinical medicine applications.

3.1 Clinical Laboratory Data

Each of the two star glyph examples in Figure 2 show twenty serum derived analyte (chemical constituent) data values each plotted on one spoke or ray of the star.

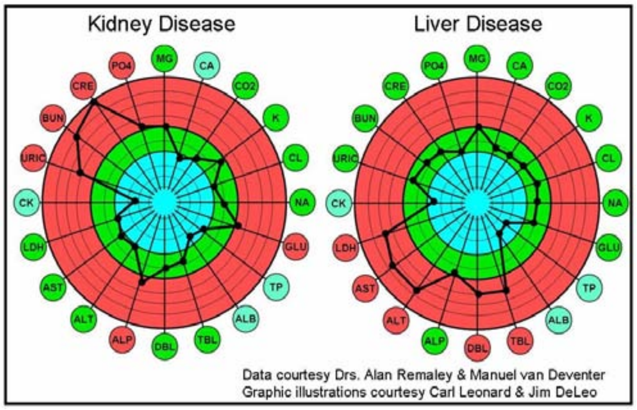


Figure 2. Star glyph plots of serum analyte values associated with a patient with kidney disease (left) and a patient with liver disease (right) with below reference values plotted in inner blue band, normal reference range values in the middle green band and high reference range values in the outer red band. The outer analyte identifier circles are color coded to indicate in which of the three bands the individual analyte values fall. (Note: rings are .1 apart.)

The serum analyte values represented in Figure 2 were derived from serum samples drawn from two different patients and produced by an clinical chemistry automated analyzer in the NIH Clinical Center Department of Laboratory Medicine and made available by Alan Remaley, MD, PhD. In this figure each analyte value is scaled on the 0-1 interval and plotted on its designated ray with 0 corresponding to the center of the plot and 1 corresponding to the end of the ray. The scaling transformation was designed to emphasize values falling outside the normal reference ranges and to compress scaled analyte values that fall within the normal reference range inside the .4 to .6 interval (the green middle band). After the scaled data values are plotted, adjacent points are connected to form a star-like pattern i.e. a “star glyph.” One obvious advantage of the star glyph is that it gives an immediate visual impression of multifactorial data – an impression that is more readily perceived and understood by the human viewer than a list of numbers on a computer screen or on a printed page. It also shows distinct patterns that are more recognizable than those obtained with more traditional data plots. For example the analyte value differences between the kidney and liver diseased patients are immediately obvious when looking at Figure 2. Thus star glyphs can be used to suggest diagnoses and classes. [8]

3.2 Sweat Patch Data

According to NIMH researchers, skin patch tests can detect abnormal levels of markers for neural and immune function in the sweat of patients with histories of depression. If confirmed, this non-invasive technique could become an easier alternative to blood tests for predicting risk for inflammatory disorders, such as metabolic syndrome, cardiovascular disease, osteoporosis, and diabetes, which often occur with depression [9]. Figure 3 shows star glyphs constructed with sweat patch data provided by NIMH researchers Esther Sternberg, MD and Marni Silverman, PhD. The data represent protein analyte values measured from sweat patches that had been worn for 24 hours by two different women, one healthy and the other diagnosed as depressed. The third smaller one in the upper middle is considered unknown. Perhaps in time clinicians could use star glyphs like this to suggest diagnoses and to recognize disease and syndrome subtypes.

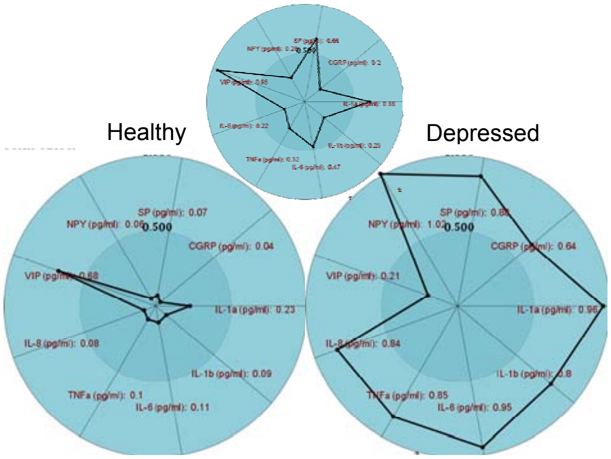


Figure 3. Star glyphs representing protein analyte value markers for neural and immune function found in the sweat of two women, one healthy (left) and one diagnosed as depressed (right).

3.3 Corticobasal Syndrome Subtyping

Corticobasal Syndrome (CBS) is a neurodegenerative disorder that has several associated major subgroups including. Alzheimer’s Disease (AD), Corticobasal Degeneration (CBD), Frontotemporal Dementia (FTD), Pick’s Disease (PD) and Progressive Supranuclear Palsy (PSP). It is very difficult to differentiate these subgroups in vivo, and currently pathological diagnosis at autopsy is the gold standard. With Jordon Grafman, PhD, NINDS we used star glyphs to plot examples of different corticobasal syndrome patients for the purpose of gaining new insights. The data were results from cognitive psychology test scores. Examples are shown in Figure 4. [10]

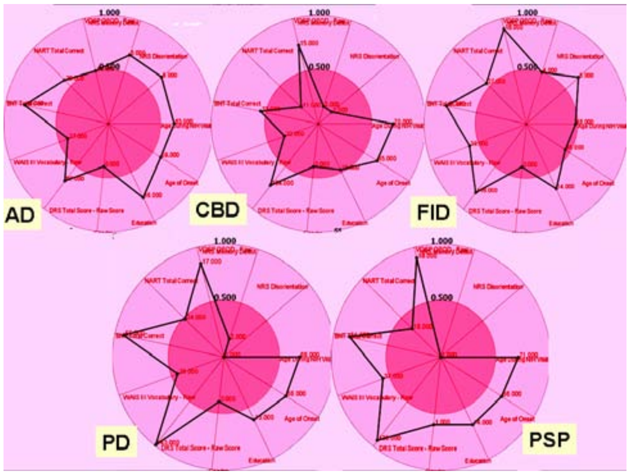


Figure 4. Star glyph plots illustrating corticobasal syndrome subtyping.

3.4 Post Traumatic Stress Disorder (PTSD)

The Beck Depression Inventory (BDI, BDI-II), created by Dr. Aaron T. Beck, is a 21-question multiple-choice (1 to 4) self-report inventory, used for measuring the severity of depression. In its present version the questionnaire is designed for individuals aged 13 and over, and is composed of items relating to symptoms of depression such as hopelessness and irritability, cognitions such as guilt or feelings of being punished, as well as physical symptoms such as fatigue, weight loss, and lack of interest in sex. We propose the use of star glyphs to visualize individual patient scores to illustrate depression as well as other psychological subtypes as illustrated with fictional BDI data in Figure 5. We have recently started a project in which we expect to apply this idea to patients suffering with war inflicted brain injury and post traumatic stress disorder. In this project we may be able to use our experience with star glyph subtyping corticobasal syndrome patients as just discussed. We are especially interested in patient psychological subtyping as well as before and after observations of the effects of holistic interventions. We plan to employ star glyph movies (discussed next) as time-varying visual records of patient wellness/illness status.

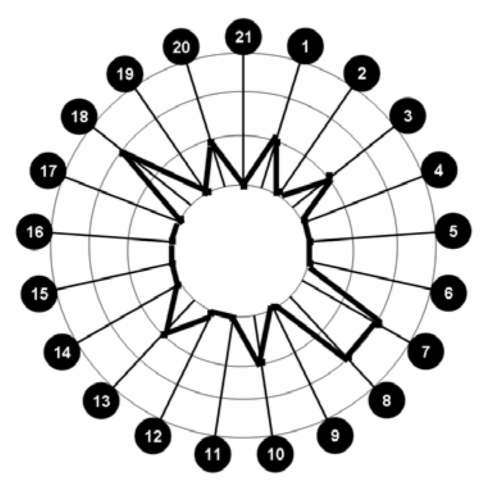


Figure 5. Star glyph with Depression Inventory data

4 STAR GLYPH MOVIES

In the examples in Section 3 above single star glyph images were used to represent static data views of related multifactorial parameter values - static because data displayed corresponds to a single time point. In most subject matter domains such as medicine, parameter values change dynamically in time. For example, a patient's serum analyte values will change over time in response to disease and treatment processes as well as just naturally in healthy states. We recently introduced the idea of creating star glyph movies to visualize such dynamic changes in data [8]. To do this, values of the parameters are collected at different points in time. Consecutive time-tagged data sets are then used to compute intermediate finer time-granularity star glyphs by linearly interpolating each of the data elements in the consecutive star glyph data values for equal time intervals. Then linear-interpolated intermediate images associated with the same subject at fixed time increments, e.g., per day, are computed. The original and interpolated images are then strung together in a time sequence and played out as a movie. Issues concerning how many data sets and how fine the granularity are application dependent and can be resolved over time with experience in creating and attempting to get knowledge from these movies. Again, scaling is an important issue here and can be optimized also with experience. We have developed software to demonstrate star glyph movies and have demonstrated it with time varying analyte values. Figure 6 shows an example of this with star glyphs representing analyte values and the time dimension indicated by means of the blue time-shadow images. Star glyph movies provide time as another dimension for knowledge data discovery. Star glyph movies could be made to illustrate serum analyte values and psychological variables such as those in the Beck Depression Inventory change over time. Side-by-side star glyph movies of patient and normal volunteer subject data could be displayed. The star glyph movies would likely demonstrate rates and magnitudes of parameter value changes over the course of compared treatments and may help to identify crucial time windows that predict treatment successes and failures. Also many diseases have flare periods followed by quiescent periods. Star glyph movies may be useful in identifying flare periods as well as the cyclical aspects of certain disease manifestations.

5 DATA SCALING

Figure 6 shows scaling differences created by star glyph movie shadow tracings. In the image on the left, the normal reference range was scaled to the .4 to .6 interval. In the one on the right the entire normal reference range was scaled to .5. There are many ways to scale data, such as range scaling, sigmoid function transformed z-score scaling and nonparametric (rank order) scaling. Scaling selection must be application specific.

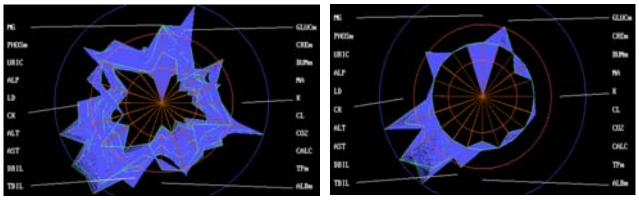


Figure 6. Shadow tracings of star glyph movies illustrating two different scaling methods.

6 PRODUCTION SYSTEM DESIGN

We would like to build the production-quality data visualization and computational intelligent system illustrated in Figure 7.

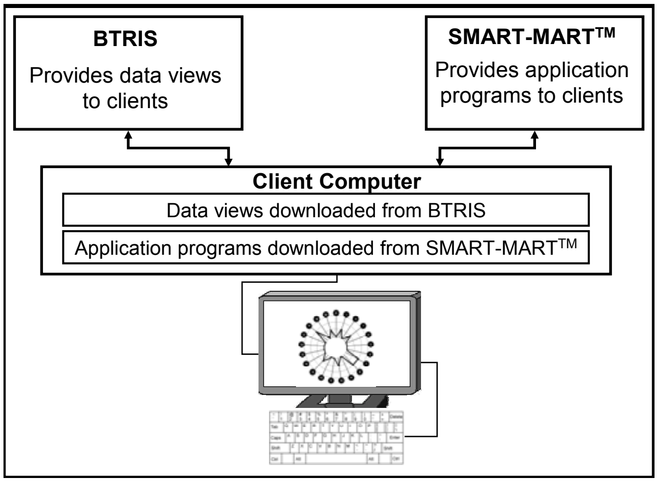

BTRIS
Provides data views to clients
SMART-MART™
Provides application programs to clients
Client Computer
Data views downloaded from BTRIS
Application programs downloaded from SMART-MART™

Figure 7. Production-quality system design for downloading data views from BTRIS and application programs from SMART-MART™

The purpose of this system is to facilitate biomedical research by making data views and application programs readily available to biomedical researchers. The design depicted in Figure 7 shows both data views and application programs downloaded into a client's computer. Having both data and application programs resident in the client's computer assures data confidentiality, which is essential in data-sensitive applications such as medicine. We refer to the application program server as SMART-MART^{TM} which we envision to be a generally available web accessible server containing a library of data visualization and other computational tools provided by contributors by means of crowdsourcing (discussed next) as an on-going process. For our work at the NIH our client principle investigators would draw data views from BTRIS and other sources. Interfaces to other data repositories could also be developed.

7 CROWDSOURCING FOR SOFTWARE

We would like to use crowdsourcing to cooperatively develop the production system just described. Crowdsourcing means to use an open call to outsource to an undefined community (the "crowd") tasks that are traditionally performed by employees and contractors. It includes contests, competitions and challenges. In his book about crowdsourcing Jeff Howe suggests that it potentially gathers the most fit with the most relevant and innovative ideas to perform the required task [11]. Crowdsourcing can yield contributions from amateurs and volunteers working in their spare time, and from professional experts and small businesses unknown to the initiating organization. Benefits of crowdsourcing may include (1) activation of intrinsic motivating incentives [12], (2) tapping a wider range of talent, (3) more heterogeneous solutions, (4) quicker solutions at no cost or low cost, (5) the crowd gets to feel brand-building kinship with the crowdsourcer and with other crowd members and (6) reward in the form of shared results may be sufficient. Our idea is that we would be the crowdsourcer that provides detailed software design documentation describing a set of data visualization tools starting with star glyphs and that we would guide the crowdsourcing process in developing production-grade software modules to be made operational in the design concept illustrated in Figure 7. We would like to use crowdsourcing in an on-going basis to continually develop and refine a library of data visualization and other computational intelligent tools to support biomedical data mining and knowledge discovery in an on-going basis. We propose starting with producing production-quality star glyph software and having it work with BTRIS-provided data as the first learning example.

8 THE AMERICA COMPETES ACT

It has been pointed out that crowdsourcing is not new. The term was first coined by Jeff Howe in a June 2006 *Wired* magazine article "The Rise of Crowdsourcing." Projects which made use of group intelligence, such as the LazyWeb or Luis von Ahn's ESP Game, predate the word "crowdsourcing" by manyl years. One thing that is new however is that it is now possible for the federal government to engage in crowdsourcing by virtue of the America Competes Act. This Act was first signed into law on August 9, 2007. Its purpose is "to invest in innovation through research and development, and to improve the competitiveness of the United States." President Obama signed a revised version of the "COMPETES Act" in January 2011. This version is specifically designed to harness America's scientific and technological ingenuity and in particular, it identifies health care improvement, better use of information technology and new product development as specific objectives. The Act gives every federal government department and agency the authority to conduct contests, competitions and challenges – methods that have demonstrated records of accomplishment for accelerating problem solving by tapping top talent and expertise. Under the Act Federal agencies may outline a problem they would like solved on Challenge.gov. At the time of this writing the NIH Office of the Director is establishing policy to ensure that NIH is compliant with the COMPETES Act. Once this policy is in place we hope to be able to advertise the crowdsourcing initiative just described. Initial announcements will be made through the NIH Biomedical Computing Interest Group (BCIG). To have your name placed on the BCIG listserver list please contact Jim DeLeo (first author) at e-mail address jdeleo@nih.gov.

9 SUMMARY

We have described our current work in applying star glyph data visualization methods to NIH clinical research data and suggested a crowdsourcing approach to develop data visualization and other data mining software compatible with data in the NIH BTRIS System. When such these methods are made available in standard, easy-to-use and readily-available software packages they are likely to become indispensible tools for gaining new insights and new knowledge in medicine as well as in other disciplines.

ACKNOWLEDGEMENTS

This work was supported in part by intramural research funds from the NIH Clinical Center and the National Library of Medicine.

REFERENCES

[1] M. Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. Proc. of Visualization '94, p. 326-33, 1994.

[2] R.J. Littlefield. Using the GLYPH concept to create user-definable display formats. Proc. NCGA '83, pp. 697-706, 1983.

[3] J.H. Siegel, E.J. Farrell, R.M. Goldwyn, H.P. Friedman, The surgical implication of physiologic patterns in myocardial infarction shock. Surgery, Vol. 72, pp. 126-141, 1972.

[4] D.F. Andrews. Plots of high dimensional data. Biometrics, Vol. 28, pp. 125-136, 1972

[5] H. Chernoff. The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association, Vol. 68, pp. 361-368, 1973.

[6] G. Grinstein, R. Pickett, M.G. Williams. EXVIS: an exploratory visualization environment. Graphics Interface '89, 1989.

[7] .J. Beddow. 'Shape Coding of Multidimensional Data on a Mircocomputer Display', Visualization '90, San Francisco, CA, 1990, pp. 238-246.

[8] J. DeLeo. Star glyphs: looking at laboratory data in a new way, ADVANCE for Administrators of the Laboratory, January 14, 2010.

[9] G. Cizza, A.H. Marques, F. Eskandari, I.C. Christie, S. Torvik, M.N. Silverman, T.M. Phillips and E.M. Sternberg. Elevated neuroimmune biomarkers in sweat patches and plasma of premenopausal women with major depressive disorder in remission: The POWER study. Biol Psychiatry. 64(10):907-911, 2008.

[10] M. Clinton, J. DeLeo, J. Grafman. Corticobasal Syndrome Patient Subtyping with HubuH^{TM} – a New Clustering Algorithm. Fifth International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics; 2008 October 3-4; Marina di Vietri Sul Mare, Salarno, Italy, CIBB IIASS; 2008.

[11] J. Howe. "Crowdsourcing: Why the power of the crowd is driving the future of business," Rivers Press, New York, 2008.

[12] D. Pink. "Drive: the surprising truth about what motivates us." Riverside Books, New York, 2009.

Visual Interactive Quality Assurance of Personalized Medicine Data and Treatment Subtype Assignment

Edward Y. Worbis*
Ohio State University, Computer Science Department

Raghu Machiraju†
Ohio State University, Computer Science Department

Christopher W. Bartlett‡
Battelle Center for Mathematical Medicine
in The Research Institute at Nationwide Childrens Hospital
and Department of Pediatrics, College of Medicine, The Ohio State University

William C. Ray§
Battelle Center for Mathematical Medicine
in The Research Institute at Nationwide Childrens Hospital
and Department of Pediatrics, College of Medicine, The Ohio State

Abstract

In this paper we describe PharmaForeCast, a new tool to improve predictions of patient clinical outcomes based on assigning their appropriate treatment subtype, which forms the basis of personalized medicine. Our prototype allows physicians to rapidly visualize not only the assignments of a conceptual black box algorithm for assigning patients to a treatment subgroup, but to also quickly assess the uncertainty of all the individual laboratory assays and other clinical information for determining the effect each of these potential errors has in determining the treatment subtype. The importance of this tool is providing physicians a way to effectively navigate the large amount of laboratory and other clinical information as to ensure the accuracy of the final subtype assignment through human quality control by an expert (the physician). Currently available tools used to visualize personal medical information, such as assays results or blood-typing, rarely allow for editing within the graphical user interface. Clustering accuracy can be significantly aided by human interaction when data points are plotted. Furthermore, current tools universally approach classification as a single-pass task, which ignores the useful information that may be gained by a clinician in an interactive analysis, in which which the clinician is able to provide expert editing of questionable data to improve the accuracy of the final subtype. This interaction loop can be repeated multiple times over a patient's lifetime. In this paper we describe the need for such a tool, which allows for highly interactive manipulations of personalized medical data using PharmaForeCast. Finally, we mention future improvements to the tool that could be applied in several subfields.

Keywords: Personalized Medicine, Visual Analytics, Quality Assurance

Index Terms: I.6.8 [Computing Methodologies]: Similiation and Modeling—Types of SimulationVisual; J.3 [Computer Applications]: Life and Medical Sciences—Medical Information Systems

1 Introduction

Since the very earliest attempts to apply computers to simplifying and automating the process of medical diagnosis it has been clear that one of the primary benefits of this approach would be the machine's ability to aggregate, compare, and act on an immeasurably larger volume of data than the human mind can assimilate. A computer can digest an entire genome's volume of data, and more, and return to the end user a summary of the state of an individual, condensed into a few variables and predictions. When applied in a medical context, such dimensional reduction can be incredibly powerful, and incredibly useful. [10, 5, 9, 3] It can predict the susceptibility of a person to disease, or identify those treatment options that are more likely, or less likely, to succeed – applications of which are at the heart of modern medicine's push for "Personalized Medicine".

Such dimensional reduction is, however, not without some peril. It is an inherently lossy process, presenting the end user with less information than was originally available, and condensing detailed and nuanced networks of observations into flat assessments of fact. This is not a defect of the approach, as it is exactly the result that is desired, however, it produces the insidious side effect that the error characteristics of the underlying data are completely disguised. The end user of such diagnoses is left with, effectively, two choices: to accept the results as valid regardless of the potential for error, or to attempt to understand and check the potential errors, which requires addressing the full dimensionality of the problem and negates many of the benefits of the original dimensional reduction. In reality, most clinical applications of this process fall somewhere between these extremes, with users attempting to address the assumed most likely sources of error, but in the end unable to universally address every potential error factor in every case. This process is insidiously problematic. While the drive towards personalized medicine accents the necessity to focus on the individual's specific disease state, rather than on the average presentation of a disease, this difficulty in dealing with potential errors results in error-checking being biased towards re-checking the most prevalent errors in "the average disease", rather than even the average presentation of a specific disease. As a result, measurements that are well understood to be frequently variable or erroneous, such as blood pressure or throat swab tests, are almost certain to receive closer or repeat attention, whether they are a relevant factor in a diagnosis or not. A measurement that is not well understood as problematic is far less likely to be reassessed, even if it is the largest contributor to the final result.

With PharmaForeCast we propose a Visual Analytics alternative to these options, that both automates and employs a subtle variation on the current state of the art, to produce improved results.

*e-mail: worbis.1@osu.edu
†e-mail: raghu@cse.ohio-state.edu
‡e-mail: Christopher.Bartlett@nationwidechildrens.org
§e-mail: ray.29@osu.edu

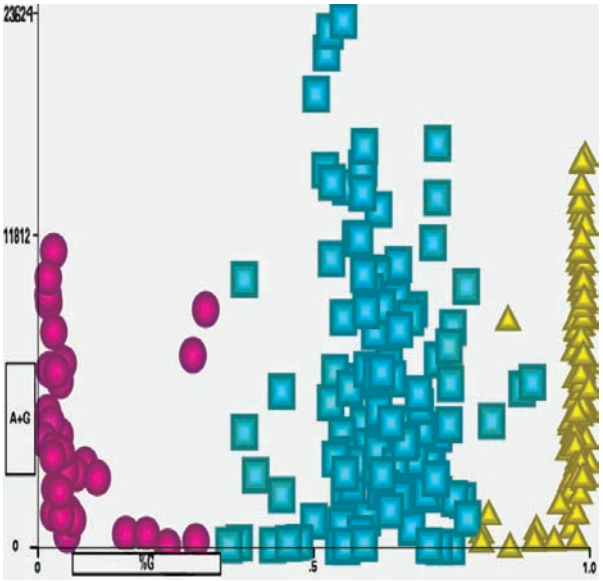


Figure 1: The PharmaForeCast prototype based on PandaSNP, a genotype editing tool, with a typical clustering result before edits. Axes correspond with strength of combined signasl along the y-direction and percentage of a base type along the x-direction. Icon/color types refer to the three allele combination types.

By engaging an iterative diagnostic updating scheme, PharmaForeCast makes certain that the human expert user is informed of the potential errors in the factors that have the largest influence on the diagnostic prediction, and that the human expert has certified that the evaluation of these factors is acceptable. This differs from the automation of the current "check the factors most likely to contain errors". Instead, by utilizing the information contained in the dimension-reducing "black box" that produces diagnoses, we can highlight the factors that were critical in producing the diagnosis, and direct the human expert's attention towards the potential errors that are most relevant to the diagnosis. If the expert updates any of these assessments, the process can be repeated and, even if the most relevant diagnostic factors change for the re-diagnosis, the user can be iteratively presented with the information necessary to critically assess the validity of the automated diagnosis.

Our prototype for this approach operates in the domain of pharmacogenomics. Pharmacogenomics derives predictions about an individual's potential drug metabolism from specific features of their genome, and uses this information to customize prescription dosages. The features of an individual's genome that might influence their metabolism of any specific drug however, are myriad, therefore significant dimensional reduction is applied in producing these predictions. Commonly, potentially numerous genomic SNP variants are used to predict the activity of several metabolic steps, which are used to identify a particular drug dosage. Several steps of dimensional reduction are applied. The practitioner attempting to apply this process is left with a suggested dosage, and perhaps some form of confidence score regarding the correctness of the score, but is left without any convenient way to assess which of the metabolic predictions might be causing reduced confidence or increased error, and without any indication of whether any of the genotype calls might be questionable, thereby inducing uncertainty. A canonical example is the prediction of warfarin metabolism, through examination of polymorphisms in the vitamin K epoxide reductase complex subunit 1 (VKORC1) and cytochrome P450 2C9 (CYP2C9) genes[8]. Because supplying a correct warfarin dose is time-critical, applying time-consuming direct-sequencing approaches to this genotyping need is problematic. Kim et. al[7] approach this problem for warfarin dose optimization by development of an "automated interpretive" application that evaluates allele-specific real-time PCR data for four VKORC1 SNPs and two CYP2C9 SNPs. These results can be produced in much less time than the direct-sequencing approach. Neither approach however, eliminates the need for the practitioner to carefully consider *all* of the possible errors in the genotyping – also a significantly time-consuming process – before acting on the suggested dose. By applying a PharmaForeCast technique, potential sources of error that are *not* important for the projected diagnosis – for example a SNP that has poor quality scores for its assignment, but is irrelevant to the metabolism at the prescribed dosage – can be eliminated from further assay, and the practitioner's attention focused on only those sources of error that can affect the prediction.

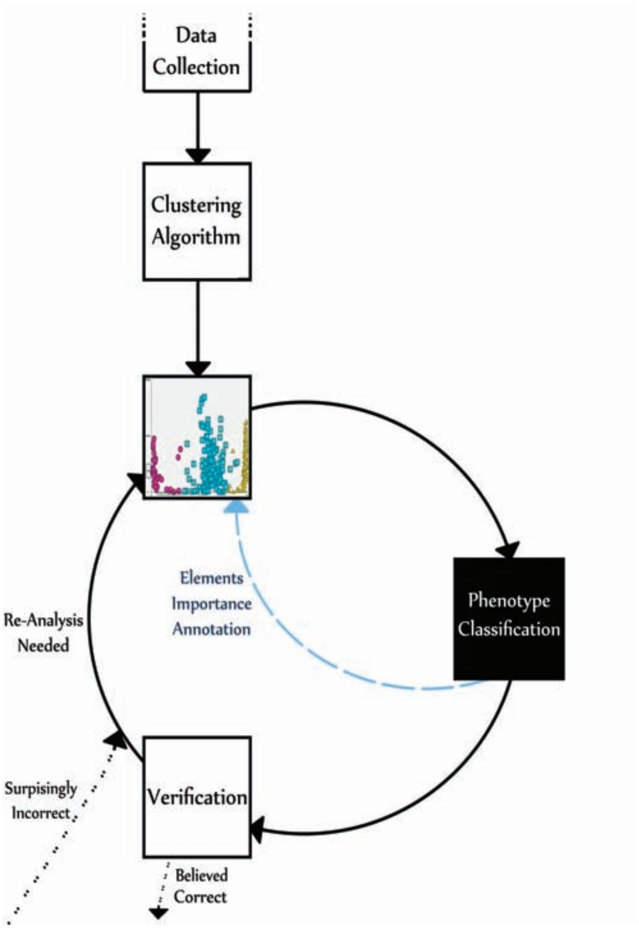


Figure 2: Visual display of PharmaForeCast's approach to data analysis. A strength of this visual analytic method is its assumption that it will be working with clinicians to readdress errors or incoporate later knowledge, while being able to work with current best clustering and classification automated methods. Black-box classification algorithms take a burden off clinicians, but at the same time PharmaForeCast takes information from them on 'most vital' data elements for aiding users by visually emphasizing which elements actually need to be reanalyzed.

2 Discussion

Our prototype application of this methodology is a Visual Analytics approach to checking, and updating SNP genotype assignments across a large number of genotypes for an individual. Because genotype assignments are not without error, and multi-locus assays of an individual's genome almost ubiquitously require the type of dimensional reduction highlighted previously, fields that use this data are in significant need of tools that allow rapid survey and identification of the salient factors that require closer examination. For pharmacogenomics, the question faced by the practitioner is "given this individual's genome, will this dosage be inadequate, adequate, or too much?". The decision must be reached by taking into consideration the patient's genotype information at anywhere from one, to dozens, to – as the field develops – eventually hundreds of different loci, many of which produce non-linear and conditional contributions to the final answer. It is virtually impossible for the practitioner to hold the complete model of these interactions in his or her head, or to hold the complete model of the possible errors. However, with as many as 3% of all genotype assignments requiring manual curation despite adequate statistical sophistication, it is a virtual requirement that the errors be addressed to maintain valid results. Our tool, PharmaForeCast, informs the user of the specific genotypes that were most influential in making the decision by providing a convenient visual survey of the quality of those assignments, and the extent to which each assignment affected the dimensionally-reduced final result. If the quality in any of the critical genotypes is unacceptable, or the assignment requires correction, the user can update the assignment and reapply the prediction to determine whether the result is the same or, if different, if any of the critical genotypes for the new result are also questionable. As shown in Figure 3, data points' shapes are made grey during reassessment if they are annotated as significant or otherwise color coded to correspond with their respective sample groups.

We believe our tool can be expanded for use in further subfields. There are many systems currently in use in medicine involving automated systems for clustering and classifying genetic and other medical data, many revolving around artificial intelligence subfields, such as evolutionary computation [12]. Unavoidably, these tools have error rates, often in excess of 5% [6, 4, 11, 14, 1, 6]. Of even greater concern is that the error rate is an average, with possibly strong inconsistentcies across sets or individual trials [2]. This means that not only are clinicians relying on incorrectly classified or clustered data, but the misclassification is inconsistent, making them harder to determine.

3 Method

The tool is meant to assist clinicians with personalized medicine through the following process:

1. The initial analysis helps in avoiding obvious misassignments or grossly ambiguous classifications.

2. The analysis data which will be used by the classifier for assigning a treatment subtype for the patient is presented visually in a way that allows the physician to drill down into the individual assays and other clinical factors to assess the quality of those assignments, any of which may have to be repeated or changed in the patient record. This is aided by the tool's GUI displaying the full data set as basic grouped and scaled items, seen in Figure 1. Data points which are categorized incorrectly stand-out strongly and the interactions for changing single or groups of points are quick and simple.

3. If a clinician has any doubts about the assignment, or if they receive information from the patient which would preclude them from being a member of the group to which they were assigned, then the traits used by the algorithm to associate the patient with the group are marked as significant.

4. The data, with marked significance, is fed back into PharmaForeCast where the data points of interest are highlighted through desaturation. This allows the doctor/user the ability to quickly reassess previously ambiguous data points, for example those which may be on the border between two clusters and had been guessed incorrectly by the clinician.

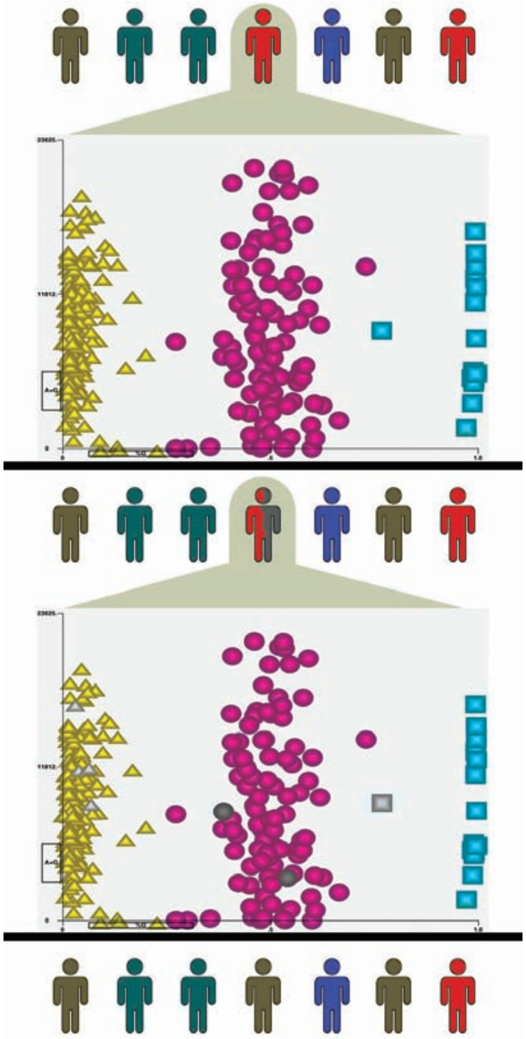


Figure 3: Dimensional reduction is a classification approach. A patient's classification (Orange) can be broken up into all of the measurements required to classify them. Measurements quality is indicated on the Y axis. If there are low-quality measurements on which the classification depends, we can update those measurements and reclassify the patient. This iterates until the patient is assigned a final classification with no low-quality, uncurated measurements involved in their classification.

5. After alterations the fixed data is run through the classification algorithm again to see if, with fixes, the patient now belongs to a similar, but different group.

6. The process is reiterated, as in the loop in Figure 2, until the clinician is satisfied, and can be resumed should new or contradictory patient information become available.

This same emphasis which allows the user to view full sets of data without being overwhelmed aids greatly during data correction. It is not practical for a user to go through thousands of individual elements which need reassignment. Our tool does allow for groups of elements to be switched collectively, but points which need changing may be scattered or mixed, and so changing each one would still present a problem in even medium scale data. The

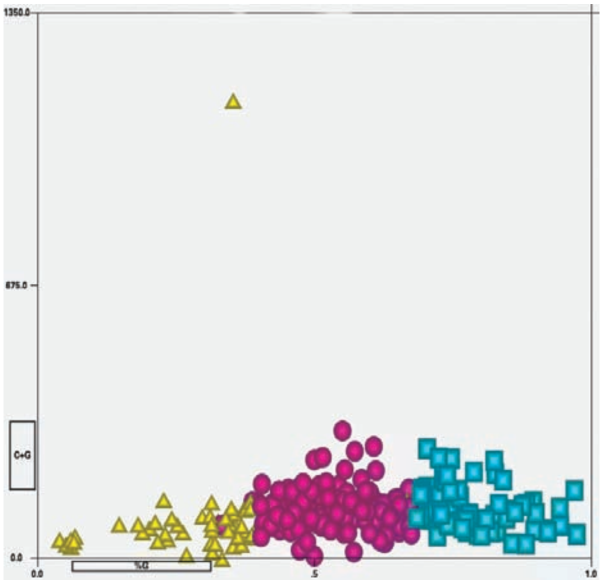


Figure 4: Despite the outlier, at top, being able to strongly influence the blackbox algorithm, the assay may lack the desired level of reliability as evidenced by the lack of clear clusters at low values on the y-axis. This datapoint could then be excluded and the algorithm rerun to check if the patient's treatment subtype was sensitive to the results on this assay. Note, the physician does not need to understand the mechanics of how this datum is used by the blackbox, but is still quite capable of performing this sensitivity analysis to understand the robustness of the assignment.

'significant' data subsets are more practical to manage. Effectively, just as the classification put the data through dimensional reduction for easy understanding, this visual technique reduces the dimensions needed for comprehending where data corrections are needed and how they should be altered.

The iterative nature of the tool emphasizes the real world issue that clinicians often learn more about a patient. Beyond the initial analysis performed patient information continues to grow. When a new symptom or analysis comes to light it is important to be able to integrate existing and novel information together as seamlessly as possible, as in Figure 3. In the best case, information collection and analysis will quickly converge to a correct diagnosis. When this does not occur, identifying whether an error occurred in old or current information, updating accurately, and reclassifying will become important. Further, like many visual analytics tools, this ability to interact with the data helps users gain a more intuitive understanding of their patient's record, and this understanding grows with each iteration.

4 Conclusion and Future Work

As we have shown, personalized medicine can benefit from our tool in many direct ways. With the possibility of doctors looking through hundreds of patients' genetic information this tool offers an advancement in throughput and accuracy for a fast approaching need.

We acknowledge a need for improvements to the tool in the future. An early enhancement we would like to add is a form of tracking past choices. As noted by Shrinivasan et al [13] and others, there is a large gain for users in seeing what has been explored in the past, especially in collaborative efforts. This would apply especially in cases where multiple doctors look at a patient'sassessment over the course of time. The ability to see what manual choices have been made before would prevent repetitive work and help gain insights into the paths of thought the previous analysts had.

Also, of use to doctors may be the ability to *show* patients, for the purpose of fuller understanding, the link between their laboratory and clinical information, and the group to which they've been assigned. For this purpose a more stylish version that still holds to an uncluttered and effective information layout may be in order.

Finally, it is forseeable that doctors and researchers may wish to compare the expressions of two patients, perhaps in a search for a common factor in a disease or because of they are family members where one has a disease with a known cause-location. Along this idea, we would like to expand the program to be able to run multiple patients' information concurrently.

References

[1] R. D. Bin and D. Risso. A novel approach to the clustering of microarray nonparametric density estimation. *BMC Bioinformatics*, 12(49), 2011.

[2] J. J. Dai, L. Lieu, and D. Rocke. Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology*, 5(1), 2006.

[3] G. S. Ginsburg and J. J. McCarthy. Personalized medicine: revolutionizing drug discovery and patient care. *Trends in Biotechnology*, 19(12):491–496, December 2001.

[4] I. M. Heid, C. Lamina, H. Kuchenhoff, G. Fischer, N. Klopp, M. Kolz, H. Grallert, C. Vollmert, S. Wagner, C. Huth, J. M. ller, M. M. ller, S. C. Hunt, A. Peters, B. Paulweber, H.-E. Wichmann, F. Kronenberg, and T. Illig. Estimating the single nucleotide polymorphism genotype misclassification from routine double measurements in large epidemiologic sample. *American Journal of Epidemiology*, 168(8):878–889, 2008.

[5] L. Hood, J. R. Heath, M. E. Phelps, and B. Lin. Systems biology and new technologies enable predictive and preventative medicine. *Science*, 306(5696):640–643, October 2004.

[6] K. I. Kim and R. Simon. Probabilistic classifiers with high-dimensional data. *Biostatistics*, 12(3):399–412, 2011.

[7] S. Kim, H. W. Lee, W. Lee, T. H. Um, C.-R. Cho, S. Chun, and W.-K. Min. New allele-specific real-time pcr system for warfarin dose genotyping equipped with an automatic interpretative function that allows rapid, accurate, and user-friendly reporting in clinical laboratories. *Thrombosis Research*, pages –, 2011.

[8] T. Klein, R. Altman, N. Eriksson, B. Gage, S. Kimmel, M. Lee, N. Limdi, D. Page, D. Roden, M. Wagner, M. Caldwell, and J. Johnson. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

[9] L. Mancinelli, M. Cronin, and W. Sade. Pharmacogenomics: The promise of personalized medicine. *The AAPS Journal*, 2(1):29–41, March 2000.

[10] R. Molidor, A. Sturn, M. Maurer, and Z. Trajanoski. New trends in bioinformatics: from genome sequence to personalized medicine. *Experimental Gerontology*, 38(10):1031–1036, October 2003.

[11] N. Pankratz. A two-stage classification approach identifies seven susceptibility gens for a simulated complex disease. In *BMC Proceedings*, 2007.

[12] A. Ramesh, C. Kambhampati, J. Monson, and P. Drew. Artificial intelligence in medicine. *Annals of The Royal College of Surgeons of England*, 88(5):334–338, September 2004.

[13] Y. Shrinivasn, D. Gotzy, and J. Lu. Connecting the dots in visual analysis. In *Proc IEEE VAST*, pages 123–130, October 2009.

[14] Z. Zhang, Y. Shi, and G. Gao. A rough set-based multiple criteria linear programming approach for the medical diagnosis and prognosis. *Expert Systems with Applications*, 36:8932–8937, 2009.

Hierarchical Summarization of Concepts for Visual Discovery Browsing – a Pilot Study

Michael J. Cairelli*, Thomas C. Rindflesch†

National Library of Medicine

Abstract

Summarization and visualization tools are needed to facilitate discovery of information in clinical research because the literature is vast and time and resources are limited. Semantic MEDLINE is a literature-based discovery browsing tool which extracts predications from the MEDLINE database and returns a graph of concepts and their relationships based on the provided search terms and parameters. These graphs can be complicated and visually cluttered, hiding valuable concepts and relationships from the user. Collapsing individual concepts into a common, generalized 'parent' concept increased readability, thereby enhancing the 'discovery opportunity'. This approach, illustrated with biomarkers for mild traumatic brain injury, can be fully automated using UMLS hierarchy.

Keywords: natural language processing, literature-based discovery, discovery browsing, semantic network, Unified Medical Language System, MEDLINE.

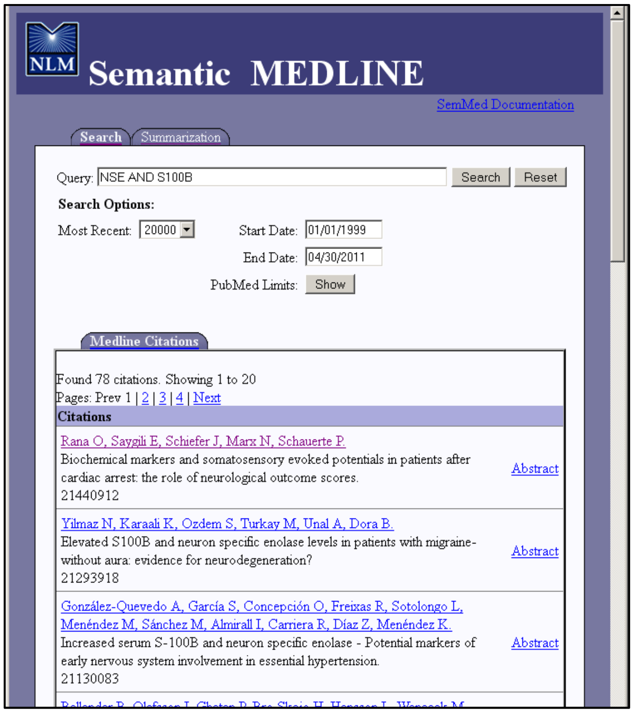
Index Terms: I.2.7 [Artificial Intelligence]: Natural Language Processing; I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods–Semantic Networks; J.3 [Life and Medical Sciences]: MEDLINE

1 Introduction to Semantic MEDLINE

Semantic MEDLINE [1] is an information management tool developed at the National Library of Medicine that assists the domain expert in searching the literature. The application exploits semantic predications from the MEDLINE database and returns a graph of concepts and their relationships summarizing retrieved text. Designed to be used by biomedical researchers and physicians, it requires no understanding of its linguistic or mathematical underpinnings. It utilizes a database created by a natural language processing system, SemRep [2], which extracts concepts and their relationships from MEDLINE citations which are mapped to the Unified Medical Language System [3] (UMLS) for consistency.

Semantic MEDLINE uses automatic summarization to isolate the most salient predications for a given search [4]. Automatic summarization allows an up-to-date and exhaustive look at the literature, unlike review papers which can be out-of-date and incomplete. This allows the time required for an exploration of the literature for a given subject to be dramatically reduced.

Most search tools provide a list ranked by date or popularity – an approach that makes little known or forgotten knowledge more difficult to find. A visual representation of concepts can provide the researcher with an accessible summary, allowing quick recognition of known and unknown relationships.



Query: NSE AND S100B

Search Options:

Most Recent: 20000

Start Date: 01/01/1999

End Date: 04/30/2011

PubMed Limits: Show

Found 78 citations. Showing 1 to 20
Pages: Prev 1 | 2 | 3 | 4 | Next

Citations

Rana O, Saygili E, Schiefer J, Marx N, Schauerte P.
Biochemical markers and somatosensory evoked potentials in patients after cardiac arrest: the role of neurological outcome scores.
21440912 Abstract

Yilmaz N, Karaali K, Ozdem S, Turkay M, Unal A, Dora B.
Elevated S100B and neuron specific enolase levels in patients with migraine-without aura: evidence for neurodegeneration?
21293918 Abstract

González-Quevedo A, García S, Concepción O, Freixas R, Sotolongo L, Menéndez M, Sánchez M, Almirall I, Carriera R, Díaz Z, Menéndez K.
Increased serum S-100B and neuron specific enolase - Potential markers of early nervous system involvement in essential hypertension.
21130083 Abstract

Figure 1. Semantic MEDLINE search tab.

1.1 Semantic MEDLINE Implementation

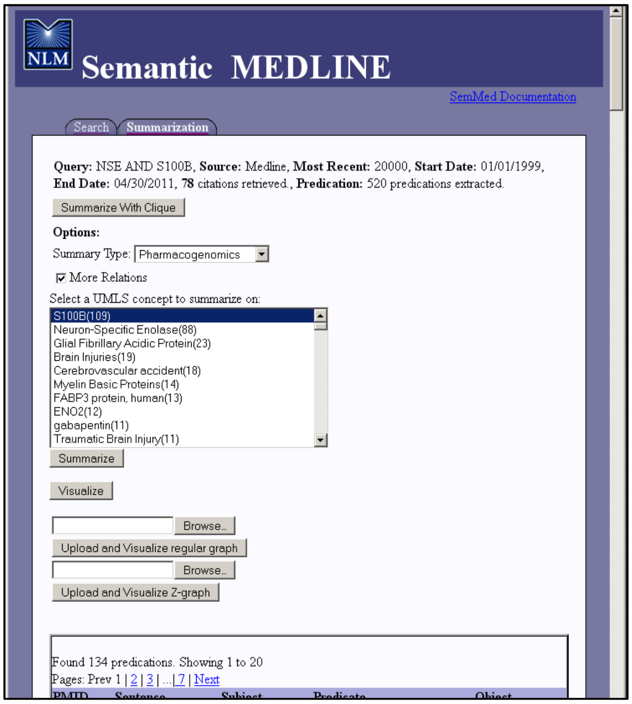
Semantic MEDLINE is a Java Enterprise Edition Web application, which utilizes open source technologies such as the Tomcat Servlet container, Apache Struts, and the Java Servlet API [1]. The Semantic MEDLINE database is a MySQL database that contains semantic predications extracted from MEDLINE citations, UMLS Metathesaurus data (such as synonyms and semantic types), and Entrez Gene data (synonyms). The database is prepopulated from plain text files generated by the SemRep application [2]. The database currently contains 27 million total predications extracted from 7.8 million total citations published between 1/2/1999 and 6/30/2011. Graph visualization is achieved using a Flash application with the Adobe Flex framework and the Flare visualization toolkit. The layouts currently used include NodeLinkTreeLayout, RadialTreeLayout, and CircleLayout.

1.2 User Interface

The user interface for Semantic MEDLINE is set up as two tabs for user input and a graph visualization for application output. The search tab allows the user to enter a specific PubMed query, select a result count limit, select a publication date range, and select other PubMed filter options. The summarization tab displays the user's search parameters, the count of citations retrieved, and the number of predications retrieved at the top.

*e-mail: mike.cairelli@nih.gov
†e-mail: trindflesch@mail.nih.gov

Lower in the form summarization options are entered including the type of summary (treatment of disease, substance interactions, diagnosis, or pharmacogenomics) and the focal node in the graph.



NLM **Semantic MEDLINE**

SemMed Documentation

Search | **Summarization**

Query: NSE AND S100B, **Source:** Medline, **Most Recent:** 20000, **Start Date:** 01/01/1999, **End Date:** 04/30/2011, **78** citations retrieved., **Predication:** 520 predications extracted.

Summarize With Clique

Options:

Summary Type: Pharmacogenomics ▾

☑ More Relations

Select a UMLS concept to summarize on:

S100B(109)
Neuron-Specific Enolase(88)
Glial Fibrillary Acidic Protein(23)
Brain Injuries(19)
Cerebrovascular accident(18)
Myelin Basic Proteins(14)
FABP3 protein, human(13)
ENO2(12)
gabapentin(11)
Traumatic Brain Injury(11)

Summarize

Visualize

Browse...

Upload and Visualize regular graph

Browse...

Upload and Visualize Z-graph

Found 134 predications. Showing 1 to 20
Pages: Prev 1 | 2 | 3 | ... | 7 | Next

PMID Sentence Subject Predicate Object

Figure 2. Semantic MEDLINE summarization tab.

Nodes are labeled and color-coded by UMLS semantic groups, such as Anatomy, Chemicals & Drugs, or Genes & Molecular Sequences (Figure 4). The relation labels for the graph are also color coded for the semantic relationship they represent (Figure 5) and are displayed with a check box for each to allow the user to decide which to include in the graph. In addition to the visualization of the graph itself, when an edge is clicked the user is provided with relationship information such as subject, object, relation, number of predications, number of citations, and a citation button which links to pop-up window with the PubMed ID linked to the PubMed entry, date of publication, title, and the abstract with the source sentence highlighted. A node search tool is also provided which shifts the graph to center on the searched concept.

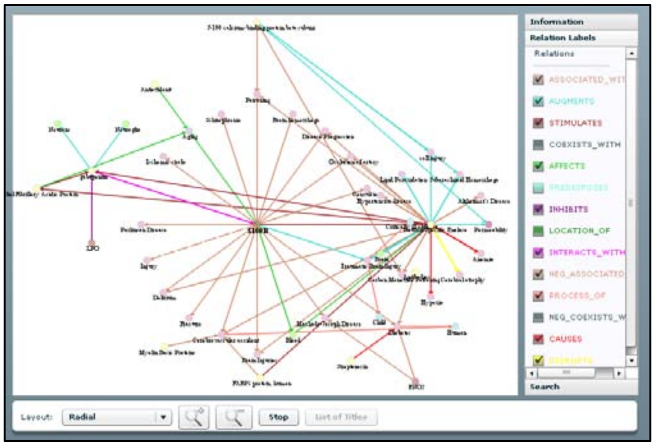


Figure 3. Semantic MEDLINE graph visualization.

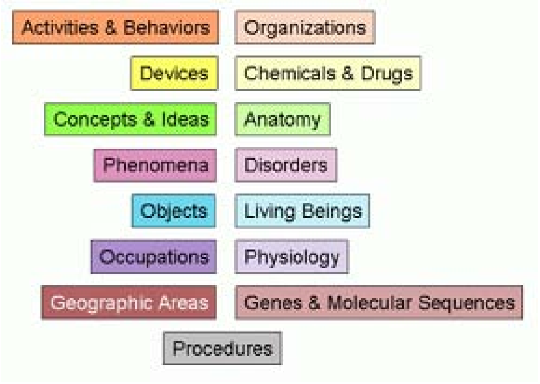
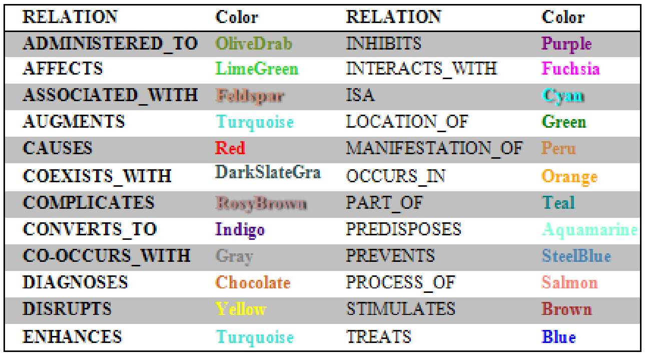


Figure 4. Node color codes by semantic type.



RELATION	Color	RELATION	Color
ADMINISTERED_TO	OliveDrab	INHIBITS	Purple
AFFECTS	LimeGreen	INTERACTS_WITH	Fuchsia
ASSOCIATED_WITH	Feldspar	ISA	Cyan
AUGMENTS	Turquoise	LOCATION_OF	Green
CAUSES	Red	MANIFESTATION_OF	Peru
COEXISTS_WITH	DarkSlateGra	OCCURS_IN	Orange
COMPLICATES	RosyBrown	PART_OF	Teal
CONVERTS_TO	Indigo	PREDISPOSES	Aquamarine
CO-OCCURS_WITH	Gray	PREVENTS	SteelBlue
DIAGNOSES	Chocolate	PROCESS_OF	Salmon
DISRUPTS	Yellow	STIMULATES	Brown
ENHANCES	Turquoise	TREATS	Blue

Figure 5. Edges color codes by relation.

1.3 Potential Design Improvements

Although Semantic MEDLINE has already contributed to significant discoveries in biomedicine (e.g. [5]), there remain opportunities for enhancing the graph visualization, thereby increasing the facility for knowledge discovery. Graphical encoding enrichments could include utilization of node size and edge thickness and multiple arcs between a pair of nodes Node size is currently constant but could be varied to represent the degree centrality (connectivity) of each node. We would also like to explore methods of displaying multiple relations between nodes, possibly using thickness of the edge to represent frequency using a logarithmic or exponential function coupled with edge bundling. Additionally, there is immediate value in providing the user the ability to merge nodes together when they belong to a common group. For example, various types of injuries could be collapsed to the common concept 'Injury'. In this paper we explore this possibility further and demonstrate the utility it garners to the user.

2 Methods

Based on previous discovery in the search for mild traumatic brain injury biomarkers, the two seed terms 'neuron-specific enolase' (NSE) and 'S100B' were used as search terms in Semantic MEDLINE with no restriction on number of citations or date of citation. Summarization was focused on substance interaction predications using the concept 'Brain Injuries' as the positional focus of the graph. Nodes that were determined by a physician to represent some concept within the UMLS semantic type 'Injury or Poisoning' were manually moved to a single position, simulating a single collapsed node.

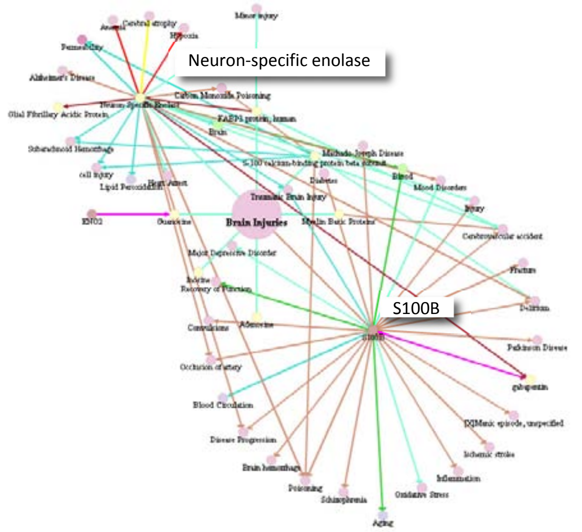


Figure 6. Semantic MEDLINE result from query search "neuron-specific enolase AND S100B".

3 RESULTS

Figure 6 shows the initial graph returned by Semantic MEDLINE. In Figure 7, the concepts that represented some form of injury are grouped together. The concept FABP3 (fatty acid binding protein 3) becomes much more evident after summarizing the injury nodes. This concept is linked by Semantic MEDLINE to a citation that identifies FABP3 as a more sensitive marker for minor brain injury than either S100B or NSE, which are two of the most studied as biomarkers of moderate to severe brain injury but not useful for mild cases (see Figure 8).

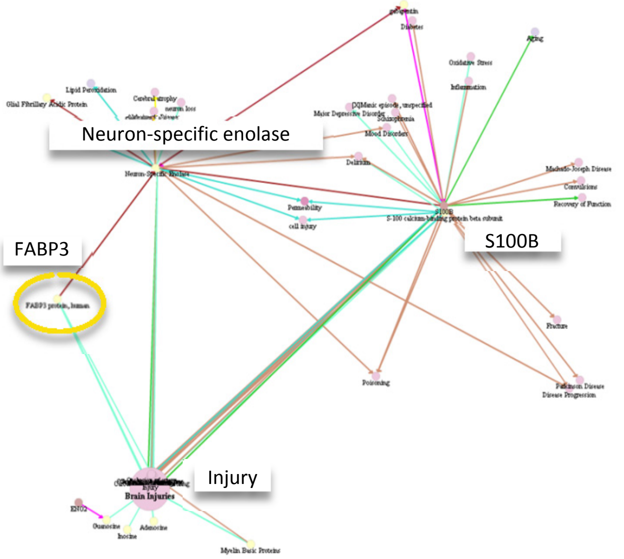
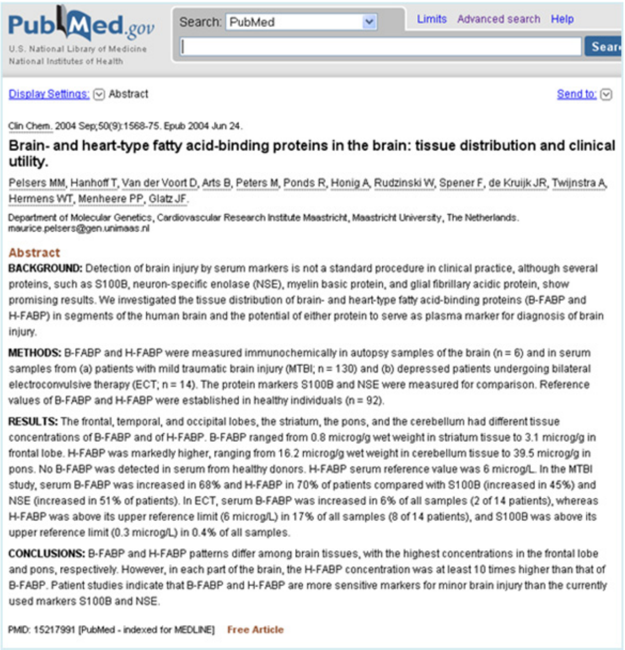


Figure 7. Nodes manually collapsed to reveal FABP3.



Clin Chem. 2004 Sep;50(9):1568-75. Epub 2004 Jun 24.

Brain- and heart-type fatty acid-binding proteins in the brain: tissue distribution and clinical utility.

Pelsers MM, Hanhoff T, Van der Voort D, Arts B, Peters M, Ponds R, Honig A, Rudzinski W, Spener F, de Kruijk JR, Twijnstra A, Hermens WT, Menheere PP, Glatz JF.

Department of Molecular Genetics, Cardiovascular Research Institute Maastricht, Maastricht University, The Netherlands.
maurice.pelsers@gen.unimaas.nl

Abstract

BACKGROUND: Detection of brain injury by serum markers is not a standard procedure in clinical practice, although several proteins, such as S100B, neuron-specific enolase (NSE), myelin basic protein, and glial fibrillary acidic protein, show promising results. We investigated the tissue distribution of brain- and heart-type fatty acid-binding proteins (B-FABP and H-FABP) in segments of the human brain and the potential of either protein to serve as plasma marker for diagnosis of brain injury.

METHODS: B-FABP and H-FABP were measured immunochemically in autopsy samples of the brain (n = 6) and in serum samples from (a) patients with mild traumatic brain injury (MTBI; n = 130) and (b) depressed patients undergoing bilateral electroconvulsive therapy (ECT; n = 14). The protein markers S100B and NSE were measured for comparison. Reference values of B-FABP and H-FABP were established in healthy individuals (n = 92).

RESULTS: The frontal, temporal, and occipital lobes, the striatum, the pons, and the cerebellum had different tissue concentrations of B-FABP and of H-FABP. B-FABP ranged from 0.8 microg/g wet weight in striatum tissue to 3.1 microg/g in frontal lobe. H-FABP was markedly higher, ranging from 16.2 microg/g wet weight in cerebellum tissue to 39.5 microg/g in pons. No B-FABP was detected in serum from healthy donors. H-FABP serum reference value was 6 microg/L. In the MTBI study, serum B-FABP was increased in 68% and H-FABP in 70% of patients compared with S100B (increased in 45%) and NSE (increased in 51% of patients). In ECT, serum B-FABP was increased in 6% of all samples (2 of 14 patients), whereas H-FABP was above its upper reference limit (6 microg/L) in 17% of all samples (8 of 14 patients), and S100B was above its upper reference limit (0.3 microg/L) in 0.4% of all samples.

CONCLUSIONS: B-FABP and H-FABP patterns differ among brain tissues, with the highest concentrations in the frontal lobe and pons, respectively. However, in each part of the brain, the H-FABP concentration was at least 10 times higher than that of B-FABP. Patient studies indicate that B-FABP and H-FABP are more sensitive markers for minor brain injury than the currently used markers S100B and NSE.

PMID: 15217991 [PubMed - indexed for MEDLINE] Free Article

Figure 8. FABP3 citation found after collapsing 'Injury' nodes.

4 DISCUSSION

Collapsing nodes with a common semantic type in complex graphs facilitates the discovery process, in which a previously unknown relation 'A–C' is discovered based on known relations 'A–B' and 'B–C'. Not only does collapsing nodes make the graph more readable by reducing clutter, but it can also show relationships between classes of concepts that may not have been obvious for specific members of the class. For example, given that concept A is a member of class I, concept B and C are members of class II, and concept C is a member of class III; if A is related to B and C is related to D, then class I is related to class III through class II (Figure 9).

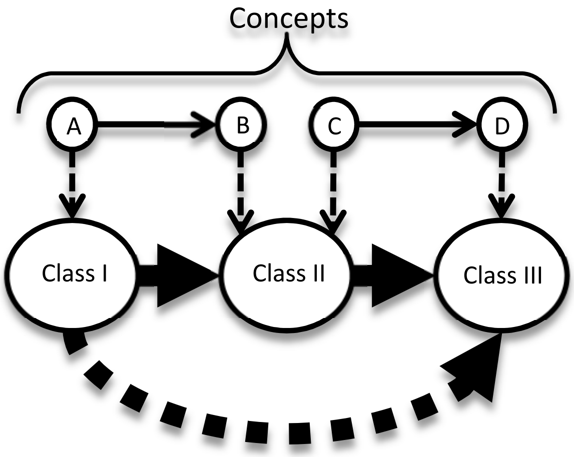


Figure 9. Nodes are generalized to their class, revealing new relationships between classes.

This method of node summarization is fully automatable. In this example we generalized to the UMLS semantic type to identify similar concepts, but this can be taken further to use the

UMLS hierarchy to find the nearest common ancestor for the nodes being collapsed.

5 Conclusion

This demonstrates the manual node summarization of a Semantic MEDLINE graph. The revised graph dramatically simplified remaining relationships, facilitating the discovery of new information (in this case, FABP3 as a substance relevant to traumatic brain injury). When automated, node summarization holds considerable potential for enhancing the discovery browsing ability of Semantic MEDLINE.

References

[1] H. Kilicoglu, M. Fiszman, A. Rodriguez, D. Shin, Anna M. Ripple, T. C. Rindflesch. Semantic MEDLINE: A Web application to manage the results of PubMed searches. *Proceedings of the Third International Symposium for Semantic Mining in Biomedicine*, (September 2008), pages 69-76, 2008.

[2] T. C. Rindflesch, M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform*. 36(6):462-77, December 2003.

[3] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 32(Database issue):D267-70, January 2004.

[4] M. Fiszman, T.C. Rindflesch, H. Kilicoglu. Abstraction summarization for managing the biomedical research literature. *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*. (Boston, May 2004), pages 76-83, 2004.

[5] C. M. Miller, T.C. Rindflesch, M. Fiszman, D. Hristovski, D. Shin, G. Rosemblat, H. Zhang, K.P. Strohl. A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. *Sleep* [in press].

Assessing Risks for Families with Inherited Cancers

Brian Drohan, Curran Kelleher, Georges Grinstein, Kevin Hughes

University of Massachusetts Lowell and Massachusetts General Hospital

ABSTRACT

A family history of disease can be a powerful clinical indicator of risk. The challenges associated with collecting a thorough family history, and maintaining that information over time, prevent the benefits of effective prophylactic treatments from reaching those with the greatest risk. We have introduced patient self-reported data entry using tablet computers, along with clinical decision support algorithms and visualizations for interpretation and analysis, into a clinical workflow capable of large scale screening and management of those at high risk for hereditary breast and ovarian cancer.

KEYWORDS: Risk Assessment, Breast Cancer, Clinical Decision Support.

INDEX TERMS: J.3 [Life and Medical Sciences]: Health—Medical Information Systems

1 INTRODUCTION

One of the great hopes of the genomic age is that by understanding and identifying genetic mutations, we can prevent disease. One of the barriers to genomic technology having a significant impact on public health has been our ability to identify those healthy individuals who are at risk of disease because they have inherited a mutation before they become sick. Until the day when everyone gets full gene sequencing as part of a screening program, we will likely rely on patient self-reported family history as signal for identifying those whom genetic testing would benefit. The difficulty this presents to our current approach is that providers are asked to collect the data in sufficient detail and then identify the wide range of syndromes, often without special training [1]. According to the Online Mendelian Inheritance in Man database of genetic disorders, there are 188 adult hereditary syndromes with at least one adult chronic disease. In this paper we discuss the example of Hereditary Breast and Ovarian Cancer (HBOC), and methods we developed to expand the identification and management of those at risk for its effects.

There are approximately 1,000,000 carriers of mutations in the genes that cause HBOC (BRCA1 and BRCA2) in the US, and of those about 50,000 (about 5%) have been identified to date [2]. We believe this poor performance is likely the highest rate of identification for any adult hereditary cancer syndrome because clinical genetic testing has been available now for thirteen years and we know a great deal about the association between the genes and treating the disease.

Most high risk women are not being identified or referred for counseling, and our risk clinics could not manage the volume if all high risk women were referred [3]. Health Information Technology (HIT) holds the key to increasing the quality of care while decreasing the cost of care [4,5,6,7]. This will be accomplished by increasing efficiency and increasing the use of Clinical Decision Support (CDS) to promulgate evidence based medical care. Thoughtful visualisations will be necessary to synthesize CDS with data from the patient and the clinician to make proper management obvious at the same time as directly supporting the clinical workflow.

We have developed a system that integrates these components into mammographic screening, genetic counseling, and surgical clinic settings.

2 LARGE SCALE METHODS

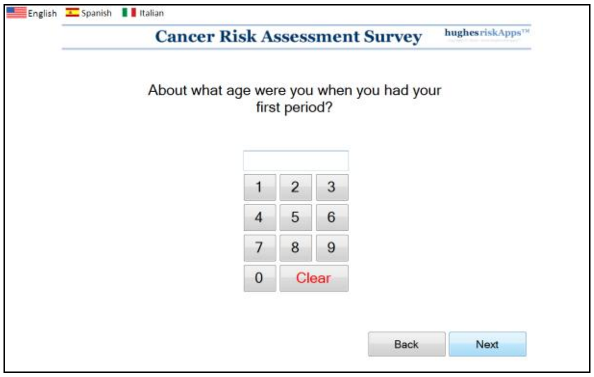
By developing an HIT infrastructure for identifying BRCA carriers, we believe the approach can scale up easily. We also expect the approach can work for many other disease areas. We also believe however it is ultimately necessary for these tools to be interoperable with other clinical systems. Unfortunately, current Electronic Health Record (EHR) systems remain digital copies of paper records, using little of the graphical or organizational power of a computer. As an example, to evaluate the risk of a hereditary condition, one must look in the demographics section, the family history section, the problem list and the lab results section to see all pieces of the puzzle. Pedigree visualizations (Figure 2) can put all this data into a single coherent picture, simplifying the clinician's work [8], but pedigrees and other visualizations remain beyond the capability of EHRs.

The need for CDS is driven by the rate in which providers are being deluged with new information. Knowledge grows exponentially, as seen in part by the dramatic rise in the number of articles in Medline and PubMed [9]. We do not believe it is reasonable to expect that providers will be able to keep up with all the information they need to manage patients. CDS provides the likely solution.

Most importantly, CDS should facilitate the best action as part of normal workflow. Today, CDS is rudimentary at best, both squandering the opportunity to increase quality and producing cynicism among providers as to its utility. EHR vendors uniformly point out alerts for drug-drug interactions and allergies as proof that they know how to accomplish CDS. In reality, these systems have failed as they do not present the information to the provider in a compelling way nor do they help the clinician follow the recommendation within the course of their normal workflow. Isaac et al identified that providers fail to act on 93.4% of drug-drug interaction alerts and 77% of allergy alerts [10].

A prototype solution

The process starts when a patient checks in and is handed a Tablet PC which displays one question per screen in a choice of languages including English, Spanish and Italian (Figure 1). Information from prior sessions pre-fills the answers to most questions, while branching logic moves over questions irrelevant to the patient. She enters risk factors, family history, and an extended review of systems.

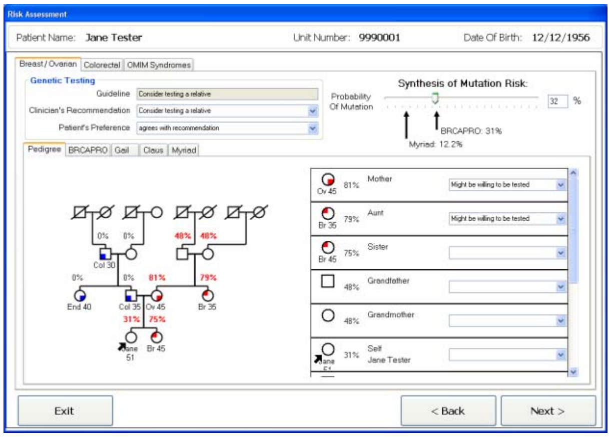


Cancer Risk Assessment Survey

About what age were you when you had your first period?

Figure 1. **A sample Hughes RiskApps patient survey question showing a simple interface with easy to read questions and options for multiple languages.**

Upon completion of the survey, risk models for breast cancer are immediately run and a summary printout is generated that displays the patient data in an intuitive form, including a pedigree. Patient information sheets, such as smoking cessation information, are generated for appropriate individuals. The staff reviews the summary printout to confirm accuracy, and makes appropriate corrections.



Patient Name: Jane Tester
Unit Number: 9990001
Date Of Birth: 12/12/1956

Genetic Testing
Guideline: Consider testing a relative
Clinician's Recommendation: Consider testing a relative
Patient's Preference: agrees with recommendation

Synthesis of Mutation Risk:
Probability Of Mutation: 32 %
BRCAPRO: 31%
Myriad: 12.2%

Ov 45	81%	Mother	Might be willing to be tested
Br 35	79%	Aunt	Might be willing to be tested
Br 45	75%	Sister	
	48%	Grandfather	
	48%	Grandmother	
Jane 51	31%	Self Jane Tester	

Figure 2. **The pedigree visualization can show risk for a whole family.**

The clinician workflow is based on an intuitive set of tabs that starts with a review of the data entered so far, progresses through the clinical encounter, and ends with all of the necessary documentation and order sheets being generated. In the risk clinic module the genetic counsellor can review the results of the risk models with the patient to help determine what the various options and likely outcomes are, and ultimately if testing is an appropriate course of action (Fig 3).

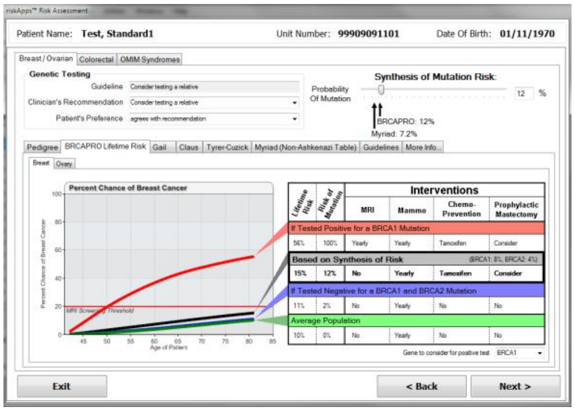
In addition to viewing the raw data, risk calculation algorithms are run and the results are displayed to the clinician using a visual representation. Graphs show results from BRCAPRO, a standard breast cancer risk assessment algorithm, run multiple times for the same family using different parameters. BRCAPRO is also run for each relevant family member, with the risk of mutation shown for each in a pedigree diagram.

Genetic testing recommendations are made within the same user interface in which the data is shown. In addition, risk of mutation is computed from another risk model, the Myriad model, and results from both BRCAPRO and Myriad are shown on a risk of mutation slider, which the clinician can set manually. Family members are listed in order of likelihood of mutation. The willingness of each to be tested can be recorded.

Lifetime risk of breast or ovarian cancer and several risk management suggestions are shown to the clinician for multiple scenarios: without testing (Current synthesis), as if the patient tested positive, as if the patient tested negative, and the population risk. Gail, Claus, MMPRO and PREMM risk model results are displayed as well. CDS suggests alternative syndromes in order of likelihood, and shows manifestations of the selected syndrome. Double clicking on a syndrome opens its page in the OMIM and Genetests Websites.

The CDS system helps the clinician find all mutation carriers by enabling the clinician to visually document the testing of family members. The tool then shows the number tested versus number of living relatives age 18 or older with a mutation risk of 10% or greater.

In the surgical module, the clinician adds details about the exam and completes collection of information using an interface tailored for this encounter (Fig 4). The immediate payback for this work is how the CDS helps the clinician develop an impression and plan. After discussion with the patient, the record is finalized and the software generates a history and physical, a letter to the referring provider, patient information sheets appropriate to the diagnosis, and a consent form if surgery is planned.



Patient Name: Test, Standard1
Unit Number: 99909091101
Date Of Birth: 01/11/1970

Genetic Testing
Guideline: Consider testing a relative
Clinician's Recommendation: Consider testing a relative
Patient's Preference: agrees with recommendation

Synthesis of Mutation Risk:
Probability Of Mutation: 12 %
BRCAPRO: 12%
Myriad: 7.2%

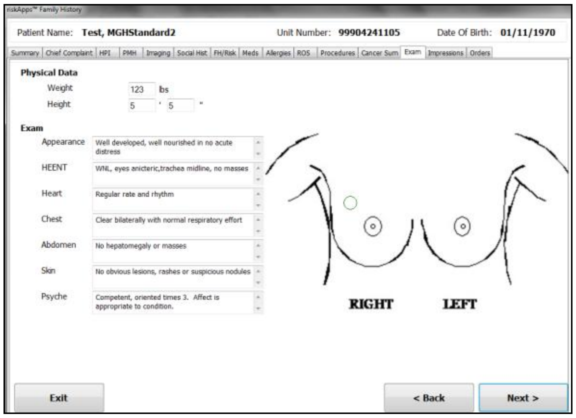
Lifetime Risk	Risk of Mutation	Interventions			
		MRI	Mammo	Chemo-Prevention	Prophylactic Mastectomy
If Tested Positive for a BRCA1 Mutation					
56%	100%	Yearly	Yearly	Tamoxifen	Consider
Based on Synthesis of Risk					(BRCA1: 8%, BRCA2: 4%)
15%	12%	No	Yearly	Tamoxifen	Consider
If Tested Negative for a BRCA1 and BRCA2 Mutation					
11%	2%	No	Yearly	No	No
Average Population					
10%	0%	No	Yearly	No	No

Gene to consider for positive test: BRCA1

Figure 3. **Clinical Decision Support guides the workflow through making a decision about genetic testing.**

On subsequent visits, pathologic results are entered which help generate additional summary documents, and generate further suggested orders (e.g., if Estrogen receptor negative, obtain Medical Oncology consultation). The approach is to keep the surgeon in line with the appropriate quality measures in real-time, rather than tell them at the end of the year how often they were not compliant.

Street Address and Electronic Mail Address



Patient Name: **Test, MGHStandard2** Unit Number: **99904241105** Date Of Birth: **01/11/1970**

Physical Data

Weight	123 lbs
Height	5 ' 5 "

Exam

Appearance	Well developed, well nourished in no acute distress
HEENT	WNL, eyes anicteric,trachea midline, no masses
Heart	Regular rate and rhythm
Chest	Clear bilaterally with normal respiratory effort
Abdomen	No hepatomegaly or masses
Skin	No obvious lesions, rashes or suspicious nodules
Psyche	Competent, oriented times 3. Affect is appropriate to condition.

Figure 4. **The data entry screen for a breast exam in the RiskApps surgical clinic module helps entering encounter-specific information.**

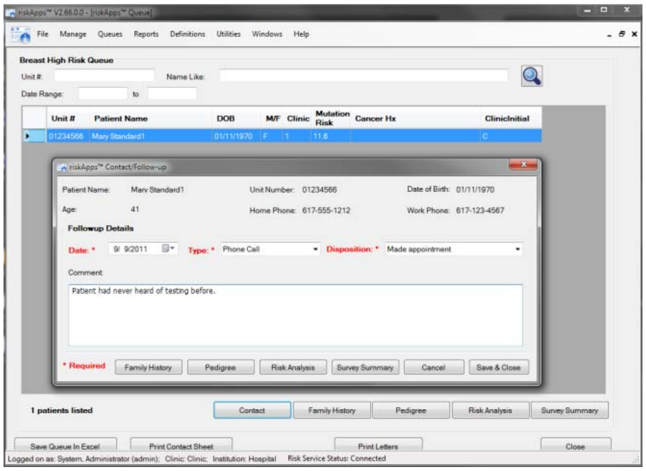
2.1 More Women Identified

Once high risk patients are identified, the next challenge is to improve the efficiency of the risk clinic to manage the influx of patients. Our challenges are to minimize clinician work, minimize redundant data entry, and minimize dictation and editing tasks.

Task	Traditional (minutes)	Our Approach (minutes)
Clinician collects family history	0 to 10	0
Data entered in risk calculator app	5 to 10	0
Data entered into pedigree drawing app	10 to 20	0
Risk level assessed	5 to 10	5
Fae to face counseling	30 to 60	30 to 60
Letters/notes generated	20 to 40	10
Total	**70 to 150**	**45 to 75**

Table 1. **Comparative time costs using the traditional approach versus the Hughes RiskApps approach.**

At the Newton Wellesley Hospital Breast Center between April of 2007 and December of 2010, 49,758 unique family histories were collected and analyzed. Of those, there were 2,255 patients whose risk of mutation were greater than ten percent and were referred for counseling. The system maintains several mechanisms for tracking those identified, including a specialized queue interface listing all at risk individuals with quick access to their family history, contact information and pending appointments at the screening center. Each identified woman is also mailed a letter that explains the risks of cancer and the testing process. This letter is copied to her primary care physician as well.



Breast High Risk Queue

Unit #: Name Like:

Date Range: to

	Unit #	Patient Name	DOB	M/F	Clinic	Mutation Risk	Cancer Hx	ClinicInitial
▶	01234566	Mary Standard1	01/11/1970	F	1	11.6		C

Patient Name: Mary Standard1 Unit Number: 01234566 Date of Birth: 01/11/1970

Age: 41 Home Phone: 617-555-1212 Work Phone: 617-123-4567

Followup Details

Date: * 9/ 9/2011 Type: * Phone Call Disposition: * Made appointment

Comment

Patient had never heard of testing before.

1 patients listed

Logged on as: System, Administrator (admin); Clinic: Clinic; Institution: Hospital Risk Service Status: Connected

Figure 5. **The high-risk patient queue gives a clinic-wide view of which patients are at the highest risk and would benefit the most from testing.**

3 DOCUMENT GENERATION

RiskApps currently generates over 80 different clinical documents specialized for a variety of workflows, saving time on dictation and cost of transcription. In the risk clinic setting, these include: a letter to the referring doctor, a letter to the patient, a progress note for the patient's chart, a letter to relatives who need testing, a letter of Medical Necessity for the patient's insurance company justifying genetic testing, and a document justifying an MRI.

The data can be easily summarized into reports required for quality measurement. The system can be set up to automatically produce performance based measures used by quality programs such as the NAPBC, NCBC, and QOPI. The system can help improve quality in 3 ways: 1.) help the clinician follow quality standards in real-time, 2.) run quality reports daily or weekly, identifying activities that do not meet the standards, or 3.) run the report yearly or when the recertification application is due.

3.1 Structured data and Standards

While EHRs do allow multiple clinicians to share data, the majority of meaningful information in the EHR is free text in encounter-based unstructured notes. As such, it cannot easily be organized, it typically cannot be used by CDS, and it is difficult to extract for research or quality initiative reporting. This emphasizes the need for structured data.

Structured data is data recorded in predefined fields (placeholders) using coding systems (ICD-9, ICD-10, CPT, SNOMED, etc.). As such, it is made machine readable. The beauty of structured data is that it allows the development of unified methods to view and interpret data. In today's EHRs, few structured data elements exist, and those that exist are mostly unpopulated. Most Clinicians will not take the time to enter structured data into an EHR because there is little return on investment.

As an example, the family history data elements needed by EHRs were published by AHIC in 2008 [11], and were these elements adopted and implemented there would be tremendous opportunity for visualizations (e.g., pedigrees) and risk algorithms. Instead, in practice the vast majority of recorded family history information is found as multiple dictated notes made by multiple clinicians, while the family history section of the EHR remains mostly ignored.

Hughes RiskApps complies with the HL7 standard for representing health records. Data from our software can be shared with any HL7 compliant software. Data can be uploaded or downloaded to any EHR that has a complete family history section and that is HL7 compliant.

4 Conclusion and Future Work

HughesRiskApps can help us realize the promise of the genomic age on a population level. As this tool is becoming more widely used, more high risk women are being identified, family history is being integrated into normal clinic workflow, more women are being cared for by risk clinics, and risk counselors are able to act with much more efficiency.

We believe the future of RiskApps, and that of all successful EHRs, will be a modular approach. Niche vendors will be able to develop approaches specific to the needs of each specialty, and then use these as frontends to any EHR [12]. In this approach, the EHR would increase its database to house common data elements, and provide the more ubiquitous functions of allergies, ePrescribing, etc. Domain specific user interfaces provide the presentation and the organization of information specific to that specialty.

References

[1] Burke W, Culver J, Pinsky L, Hall S, Reynolds SE, Yasui Y, Press N. Genetic assessment of breast cancer risk in primary care practice. *Am J Med Genet A*. 2009 Mar;149A(3):349-56.

[2] Brian Drohan, Ph.D., Constance A. Roche, MSN, RN, James C. Cusack, Jr., M.D., and Kevin S. Hughes, M.D. Hereditary Breast and Ovarian Cancer and Other Hereditary Syndromes: Using Technology to Identify Carriers.

[3] Ozanne EM, Loberg A, Hughes S, Lawrence C, Drohan B, Semine A, Jellinek M, Cronin C, Milham F, Dowd D, Block C, Lockhart D, Sharko J, Grinstein G, Hughes KS. Identification and management of women at high risk for hereditary breast/ovarian cancer syndrome. *Breast Journal*, 15(2):155-162, 2009.

[4] Melinda Beeuwkes Buntin, Matthew F. Burke, Michael C. Hoaglin and David Blumenthal. The Benefits Of Health Information Technology: A Review Of The Recent Literature Shows Predominantly Positive Results .*Health Affairs*, 30, no.3 (2011):464-471

[5] Blumenthal D, Glaser JP. Information technology comes to medicine. *N Engl J Med*. 2007;356(24):2527-2534.

[6] Committee on Quality of Health Care in America, Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: *National Academy Press*; 2001.

[7] Hillestad R, Bigelow J, Bower A, et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Affairs* (Millwood). 2005;24(5):1103-1117.

[8] Drohan B, Ozanne EM, Hughes KS. Electronic health records and the management of women at high risk of hereditary breast and ovarian cancer. *Breast Journal*. 2009 Sep-Oct;15 Suppl 1:S46-55. Review.

[9] Yoo et al. A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. *BMC Bioinformatics* 2007 8(Suppl 9):S4.

[10] Thomas Isaac, MD, MBA, MPH; Joel S. Weissman, PhD; Roger B. Davis, ScD; Michael Massagli, PhD;Adrienne Cyrulik, MPH; Daniel Z. Sands, MD, MPH; Saul N. Weingart, MD, PhD. Overrides of Medication Alerts in Ambulatory Care. *Arch Intern Med*. 2009;169(3):305-311

[11] Feero WG, Bigley MB, Brinner KM. New Standards and Enhanced Utility for Family Health History Information in the Electronic Health Record: An Update from the American Health Information Community's Family Health History Multi-Stakeholder Workgroup. *J Am Med Inform Assoc*. 2008;15:723–728.

[12] Mandl KD, Kohane IS. No small change for the health information economy. *N Engl J Med*. 2009 Mar 26;360(13):1278-81

Interactive Visualization for Understanding and Analysing Medical Data

Samar Al-Hajj*, MSc
School of Interactive Arts and Technology
Simon Fraser University

Richard Arias**, PhD
SCIENCE Lab
Simon Fraser University

Brian Fisher***, PhD
SCIENCE Lab
Simon Fraser University

Abstract

Massive amounts of biomedical data generated by the latest high throughput technologies are challenging to analyze. Visual Analytics (VA) tools and techniques are intended to amplify medical researchers' cognitive and perceptual capabilities and enable them to understand complex biomedical data. In this study, we explore how visualization tools can facilitate the exploratory analysis of this data. In order to assess and evaluate the effectiveness and usefulness of using visualization tools to enhance medical analysts' data exploration, we analyzed the use of Tableau and iPCA by biomedical researchers to explore immunological data. Our findings reveal that VA tools are efficient and powerful tools that can be integrated into healthcare systems to help health researchers get insights and generate knowledge from their complex medical data.

Keywords: iPCA, Tableau Software, Interactive Visualization.

1 Introduction

The latest high-throughput biomedical technologies used in flow cytometry produce massive amounts of medical data. The magnitude and complexity of these data are overwhelming to immunological researchers including immunologists and biologists. Analysing and extracting useful information from these data impose a great challenge on the medical research community. It is our argument that efficient and effective visualization tools can facilitate the exploration and analysis of complex biomedical data. Interactive visualizations provide biomedical researchers and analysts with efficient tools and techniques to amplify their cognitive skills and enhance their initial understanding of the data during the exploratory analysis process.

Visual Analytics (VA) is defined as "the science of analytical reasoning facilitated by interactive visual interfaces" [5]. These interactive visual interfaces rely on advanced visualizations of data and interactive techniques to accelerate the data analysis process, derive insights, acquire knowledge and optimize decision-making [6]. The implementation of interactive visualization tools was introduced in various medical disciplines to amplify analysts' cognitive capabilities and address the challenge of extracting useful information from massive datasets.

In this study, we present a case study of immunologists and biologists analyzing massive and multi-dimensional datasets using two visualization tools: iPCA (interactive Principal Component Analysis) and Tableau Software. Furthermore, we demonstrate how the integration of real-time visualization tools can help biomedical researchers uncover hidden trends in complex data and expose data patterns that are not noticeable otherwise, and ultimately facilitate the exploratory data analysis process. Finally, we show how immunologists exploited these visualization results to generate valuable qualitative information and drive new research questions.

2 Task, Materials and Data

2.1 Task and Data

In order to assess the accuracy and effectiveness of VA tools for medical data analysis, we used analysis immunological data as a case study.

Sub-Saharan Africa has the largest HIV-infected population in the world [2]. The vast majority of infants born to HIV positive mothers are not infected themselves. However, those **H**IV **E**xposed but **U**ninfected (**HEU**) infants are at a high risk of mortality during their first year of life; they suffer severe immune system deficiencies and an abnormal susceptibility to infections and diseases [7]. The causes of this mortality and morbidity are unclear and are currently the subject of a biomedical research carried out by immunologists and biologists at the Child and Family Research Institute (CFRI) in Vancouver, BC. The main analytical goal of this research project is to understand the immune responses of HEU infants and link these responses to causes of high mortality and morbidity.

The HEU dataset included laboratory data generated by Flow Cytometry Luminex high-throughput technologies at CFRI. Blood and tissue samples from HEU infants, HIV positive infants, and unexposed infants (EU) were stimulated with several infectious agents and fed into the Flow Cytometry device to measure infants' immunological responses by focusing on cytokine levels. The datasets were multidimensional, heterogeneous and complex. They included the flow cytometry data on cytokine responses to infectious agents, as well as the infants' demographics, feeding methods, and vaccine reactions data.

2.2 The Analytic Setting: Paired Analytics

To focus on the accuracy and effectiveness of VA tools in supporting the analysis of these multidimensional biomedical data, rather than wasting immunologists' time in tool training, we decided to follow a pair analytics protocol for collaborative visual analysis [8].

In a pair analytics protocol, a visual analytic Tool Expert (TE) is paired with a Subject Matter Expert (SME) to conduct a collaborative visual analytic session organized around a well-defined task, a dataset, and a visual analytic tool [8]. Since the TE lacked biomedical expertise to conduct a meaningful analysis of

* samara@sfu.ca
**ariasher@sfu.ca
***bfisher@sfu.ca

the HEU data, and the SME lacked tool expertise to operate the visual analytic tool proficiently, their collaboration was required to make the most of the visual analytic sessions. In our case, the SMEs were biologists and immunologists. The TE was the main author of this paper. The pair analysis was structured to help immunologists exploit the VA tool and increase the speed, efficiency, and accuracy of the exploratory data analysis process [8]. Both experts worked together and exchanged expertise to understand the HEU data and assess the relevance of using two different visualization tools (iPCA and Tableau) for exploring biomedical HEU data.

3 Visualizations And Insights

Intuitive and interactive data visualizations facilitated the exploratory analysis of HEU data and enabled immunologists and biomedical researchers to analyse and interact with HEU data at various levels of abstractions to identify trends, patterns and formulate hypotheses.

To study HEU infants' immune system reactions to infectious agents, we explored HEU and EU infants' cytokine reactions using two interactive visualization tools: iPCA and Tableau Software.

3.1 Interactive Principal Component Analysis (iPCA)

Interactive Principal Component Analysis (iPCA) is an interactive visual analysis tool developed by the Charlotte Visualization Centre. iPCA uses the Principal Component Analysis (PCA) technique to reduce high dimensional datasets and convert data into new meaningful representations in order to facilitate users analytical reasoning and expedite the data exploratory analysis process [3]. Since HEU datasets were multidimensional, we plotted the HEU data in iPCA to visualize the reduction of variables representing infants' cytokine reactions into principal components, and to analyze the distribution and contribution of variables to the principal components.

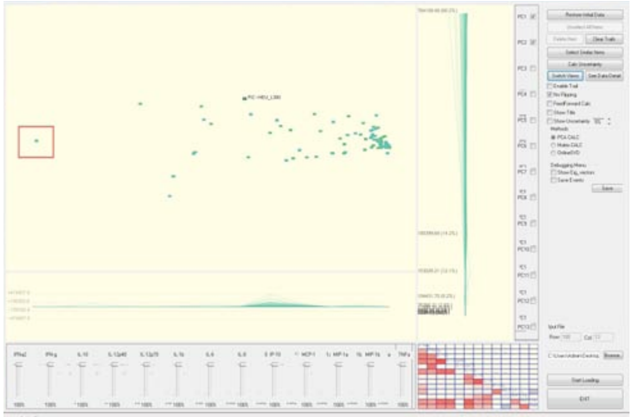


Figure 1. Visualization of Immunological data in iPCA Views: The Projection View, the Eigenvector View, the Data View and the Correlation View.

Fig. 1 shows cytokine reactions to one treatment (pIC). Each colour represents one group (i.e. HUE or EU), and each dot represents the values for one patient in that group. iPCA visually reveals the relationship between data variables, highlights outliers and provides immunologists with a comprehensive overview of existing correlations among pairs of variables. In Fig. 1, for example, outliers are quickly detected on the left side (highlighted here by the red box). It is also evident that the first principal component accounts for most of the variability in this dataset (60.2%). The slides associated with each variable (i.e. cytokines) allowed the TE to quickly show SME each variable' unique contribute to the principal components. In this particular visualization, most variables did not contribute significantly to the constitution of the first two principal components. One exception was the variable representing the cytokine IP10. Figure 2, shows the state of the visualization after the TE interacts with the "IP-10" slide dropping its contribution to the principal components to zero (highlighted by the red box). It was visually evident the dramatic reorganization of the values on the scatter plot representing the first two principal components. The first principal component, for example dropped from accounting for 60.2% to accounting for 42.3% of the variability in the dataset, while the second principal component increased from 14.2 % to 26.8%.

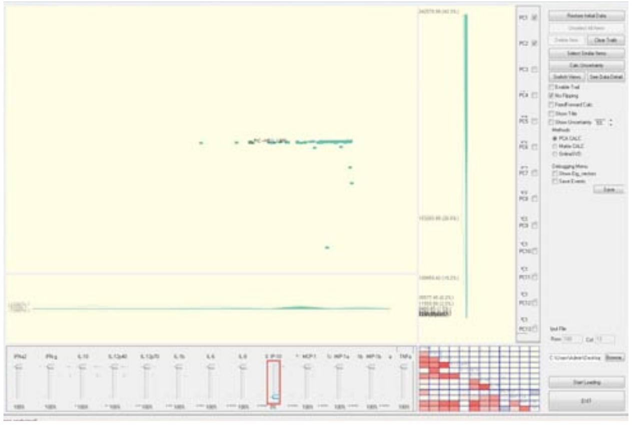


Figure 2. Interaction with individual variables to visualize their impact on the constitution of principal components

According to this analysis, the cytokine IP10 seemed to be a good candidate for further statistical analysis. To verify the analysis outcome, iPCA offered a matrix of correlations of pairs of variables. Every variable in the matrix is plotted against other existing variables to determine correlation coefficients (See Fig. 3). The correlation matrix proved very useful to quickly confirm the independence between IP10 and all the other cytokines in terms of responding to pIC treatment, which is visualized by the absence of dark red colours (i.e. indicators of high correlation) on the row corresponding to IP10 (highlighted by the red box).

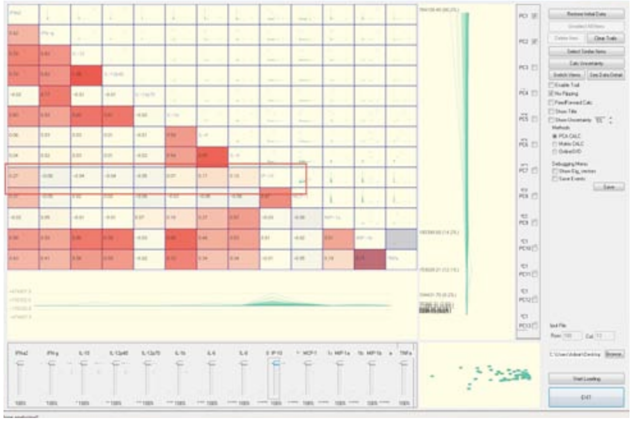


Figure 3. Correlation matrix

In summary, immunologists were able to quickly explore the HEU data and interactively identify which variables were more and less relevant for further statistical analyses. iPCA enabled immunologists to interact with the HEU data in real-time. Each time they changed a data item in one view; the change was automatically reflected in other views, giving immunologists the ability to understand data patterns and characteristics. By interacting with data, immunologists understood the influence and intuitively perceived the weight of separate variables on the constitution of the principal components. iPCA also allowed immunologists to visually detect and investigate outliers (Fig. 1) and their corresponding data items by eliminating an outlier from the data and observing its effect on the overall data visualization.

3.2 Tableau Software

Another visualization tool was used to visualize the HEU datasets: Tableau Software. Tableau is a commercial tool used for data exploration; it uses interactive visual dashboards to represent data and facilitate the exploratory data analysis process [1]. In order to compare the HEU, HIV and EU infants' cytokine reactions to infectious agents, we plotted and visually compared infants' IP10 cytokine reactions to each one of the 6 types of stimulations: CpG, pIC, R848, LPS, PG, and PAM, as well as the unstimulated control: Unstim.

The outcome of the graph, as shown in Fig. 4, depicts the infants' average cytokine reactions. Tableau represents cytokine reactions with different colors and saturations to reveal trends and show patterns in data. These patterns reflect variation across infants' groups, indicating that cytokine reactions are cohort-specific and vary between HIV unexposed and exposed infants. Tableau enables immunologists to drill down the HEU datasets and get further individual detailed information. Each bar of the graph represents all types of cytokine reactions per infant. The value of each cytokine reaction dictates the height of the bar. The shape of the bars represents a powerful visualization that provides immunologists with a comprehensive picture of the difference among HEU, HIV and EU cytokine reactions.

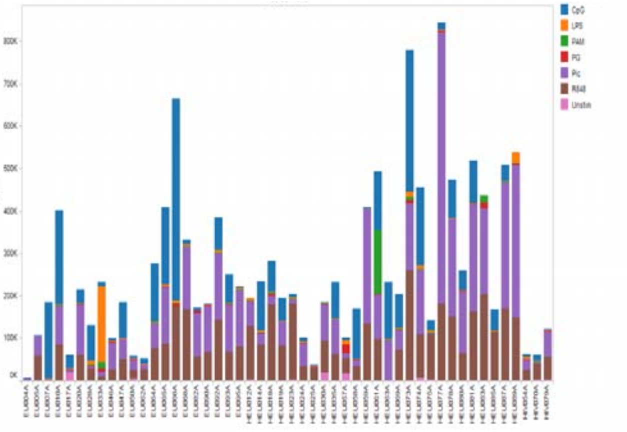


Figure 4. EU, HEU, and HIV infants' cytokine reactions to stimulations. [9]

The interactivity aspect of the HEU Tableau visualization supports immunologists' visual data exploration; it enables immunologists to hover the mouse over any particular patient and get on-demand detailed accurate statistical information about the infant's cytokine reactions to each of the stimulations. The observed variations in cytokine reactions proved the immunologists' hypothesis stating that HIV exposed but uninfected infants experience less immune defense against infectious agents compared to unexposed infants.

Immunologists were able to observe how HEU infants reacted differently to the majority of the stimulations, indicating that HEU infants' immune system differ from unexposed infants' immune system in terms of reacting to invading infectious agents and susceptibility to disease.

4 Discussion

Visual Analytics tools and techniques amplify biomedical analysts' cognitive and perceptual skills in order to observe and comprehend complex medical data, derive scientific insights and acquire knowledge to accelerate health discoveries [8]. Through real-time interactive visualization, Visual Analytics empowers biomedical analysts with the ability to reason and make sense of data under investigation.

Immunologists expressed their design and features preferences when interacting with the iPCA and Tableau visualization tools, which could be pivotal to inform the redesign of current tools to better fit the exploratory data analysis process. On one hand, immunologists pointed out flaws when interacting with Tableau. Immunologists were mainly concerned with the lack of the correlation option in Tableau, a feature that is perceived to be beneficial to the data exploration process. On the other hand, during the pair analysis session, the SMEs reported difficulties when interacting with iPCA. Something expected since iPCA is an experimental VA tool. Firstly, iPCA needs filtering; selecting and deleting groups as form of filtering is a cumbersome process. Secondly, SMEs couldn't directly access raw data from iPCA as the tool does not offer this functionality. The TE had to open a spreadsheet with the raw values on a second screen in order to have simultaneous access to the raw data. Thirdly, iPCA automatically assigned colours to groups and did not offer options for colour customizing to make differences between groups more visually salient. Fourthly, iPCA did not offer quantitative information about the exact contribution of each variable to each principal component. This information needed to be deduced by interacting with every single slide. A table with values may prove to be a better, faster, and more precise way to reach similar conclusions. Finally, iPCA did not provide features to export data. Since principal component analysis is an intermediary process in the statistical analysis process of multidimensional data, iPCA should enable users to export data to a statistical package to determine whether there is any statistical difference among the groups.

iPCA and Tableau encounter few pitfalls that constraint and limit their applications to our current HEU data. However, our preliminary findings indicate that VA tools support biomedical data exploration and knowledge dissemination. iPCA and Tableau visualization examples validate the relevance of using efficient VA tools and techniques for healthcare applications. iPCA and Tableau visualizations reveal important features about the HEU dataset and illustrate the useful application of Visual Analytics for data exploratory analysis. Furthermore, VA promoted collaboration and dissemination of information among health professionals, which is vital for the decision-making process [4]. Tableau visualization software enabled biomedical researchers to disseminate, share and communicate analysis results with a variety of audience through the creation of dashboards. Produced dashboards can be published to communicate information, interactively explore results and disseminate knowledge to colleagues as well as patients to ease dialogue with them. This is an efficient way to share knowledge and promote collaborative analytical reasoning.

Based on the visualization of the HEU data, VA provided powerful interactive visualizations needed to assist immunologists and medical researchers in data exploration as well as to generate hypotheses and test these hypotheses. VA enabled immunologists to engage and interact with the high dimensional HEU data, discover details and relationships among data variables, recognize relevant patterns, identify data clusters and outliers, and ultimately advance their research. Immunologists' experience motivates other health professionals and promotes the use of VA tools and techniques to explore complex data and to integrate powerful and effective visualization software in clinical practices.

5 Conclusion

High throughput flow cytometry technology provides immunologists with complex and multifaceted data. Exploring and examining massive and unstructured medical data exceed the ability of health professionals to synthesize meaningful information. Interactive and dynamic graphical presentations of data empower immunologists with a better perception of the HIV disease progression and a good understanding of the HEU infants' immunodeficiency. Visual Analytics uses interactive and intuitive visualizations to help medical researchers determine hypotheses, formulate research questions and conduct exploratory data analysis efficiently. Effective visualization of the HEU data represents a fundamental step in the data analysis process that can guide relevant medical discoveries and gain insights into valuable medical information. Understanding complex HEU data and drawing valid conclusions enable immunologists to identify the health determinants of HEU infants and eventually make decisive public health interventions to reduce HEU infants' sufferings and bring changes to the lives of over 300,000 HEU infants born annually [7].

We identified emerging challenges with iPCA and Tableau that could provide opportunities to improve the current version of the tools or design new tools that accommodate the needs of biomedical researchers and analysts. Further research into the potential implementation of visualization software for medical applications will determine how these visualizations can significantly affect the way analyst look at their data and guide effective integration of VA techniques and tools in various health care systems to help medical researchers generate knowledge and gain insights.

6 Acknowledgement

The authors would like to thank immunologists and biomedical researchers at the Child and Family Research Centre (CFRI) as well as our colleagues at SFU SCIENCE Lab for their collaboration and support.

References

[1] Tableau Software, 2011. http://www.tableausoftware.com/

[2] United Nations Programs on HIV/AIDS (UNAIDS). Report on the global HIV/AIDS Epidemic. Global report. 2010

[3] D. H Jeong et al., “iPCA: An Interactive System for PCA-based Visual Analytics,” in *Computer Graphics Forum*, vol. 28, 2009, 767–77 http://www.vrissue.com/portfolio/iPCA/iPCA.php

[4] D. Keim et al., “Visual analytics: Scope and challenges,” *Visual Data Mining* (2008): 76–90.

[5] J. J Thomas and K. A Cook, “Illuminating the path: The research and development agenda for visual analytics,” *IEEE Computer Society* (2005).

[6] P. N Johnson-Laird, *How we reason* (Oxford University Press, USA, 2006).

[7] Peter Wall Institute for Advanced Studies. “HIV-Exposed but Uninfected (HEU) Infants: Exploration of the Causes of Enhanced Morbidity and Mortality”. The University of British Columbia. 2009.

[8] R. Arias-Hernandez, L. T. Kaastra, T. M. Green, and B. Fisher, “Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics,” in System Sciences (HICSS), 2011 44th Hawaii International Conference on, 2011, p. 1-10.

[9] S. Al-Hajj, E. Fortuno III, and B. Fisher, “Data Visualization of Immunological Competence of HIV Exposed but Uninfected (HEU) infants”, *IEEE Visual Analytics for Science and Technology (VAST)*, 2011.

Quantitating pathogenic biofilm architecture in biopsied tissue

Shareef M. Dabdoub*
The Ohio State University Biophysics Program

Brian A. VanderBrink
Nationwide Children's

Sheryl S. Justice
Center for Microbial Pathogenesis at
Nationwide Children's Research Institute

William C. Ray†
Battelle Center for Mathematical Medicine at
Nationwide Children's Research Institute

Abstract

Accurate diagnosis and treatment of biofilm infections require identification of the pathogenic organism(s) as well as determining the progress of the disease. Current tools in clinical use, including culturing and PCR tests, are extremely useful for identifying organisms, but are destructive in nature – resulting in the loss of important information regarding biofilm architecture and state. Improving clinical understanding of these, often treatment-resistant, infections is of great importance, and new non-destructive imaging-based tools must be developed in order to gather crucial information regarding disease.

Here we present new software, *ProkaryMetrics*, designed to take advantage of available microscopy imaging modalities, providing a unique platform for 3D imaging and analysis of biofilm samples. We demonstrate the software capabilities by analysis of murine tissue biopsy samples containing uropathogenic *Escherichia coli* biofilms: wild type UTI89 and UTI89Δ*kpsF* strains. Using *ProkaryMetrics*, we establish significant architectural differences with qualitative 3D visualizations as well as quantitative measurements including volumetric biofilm size, bacterial counts, community density, orientations, and lengths.

Index Terms: J.3 [Computer Applications]: Life and Medical Science—Biology and genetics; D.2.11 [Software]: Software Architectures—Domain-specific architectures

1 Introduction

Proper treatment of an infectious disease requires the identification of the causative agent and state of the disease, as well as the susceptibility of the organism to standard treatment. While culture-based techniques remain the gold standard for identification of bacterial and fungal pathogens, diagnostically significant features of an infection also include aspects of the current activity and state of the pathogen, in addition to its simple identity as available through culturing [1]. There is increasing recognition that the state information lost through the culturing process can be critical for properly identifying causative agents and appropriate treatment. However, there is a dearth of quantitative approaches to acquiring state-related measures such as pathogen morphology, biofilm/community organization, and cellular localization from pathology specimens. The current approaches rely either on automated applications of computer-vision, or on manual applications of expert-user visual assessments from (typically) serial microscopy/histology sections. Unfortunately, there are significant impediments to both of these approaches, as, at the diagnostic endpoint there is insufficient homogeneity across either samples or image-acquisition systems for any automated system to be universally, or even widely successful, and simultaneously there is sufficient variation in clinical-user expertise that evaluations from different experts are not quantitatively comparable. Until significantly more sophisticated imaging capabilities are routinely available to endpoint clinical caregivers, any successful approach to integrating quantitative assessments of pathogen state information into treatment decisions, will require systems that can extract quantitatively comparable data from numerous disparate imaging systems and imaging modalities, without requiring more than a lay expertise in applying or adapting the computational approach. Explicitly, we propose that enabling a rural physician with a white-light microscope to, with human intervention, make quantitatively comparable measurements of clinically relevant variables, to those produced by a research laboratory with a scanning confocal instrument, is far more clinically useful than developing an automated approach for the confocal data alone.

To this end, we introduce *ProkaryMetrics*, a Visual Analytics tool for extracting quantitative measures of pathogen community morphology, density and architecture from microscopy images. *ProkaryMetrics* leverages straightforward computer vision and volume segmentation/visualization approaches that can be applied on commodity hardware, to provide a guided interface through which a human expert can rapidly annotate salient pathogen/community features for quantitative analysis. By applying algorithmic volume segmentation/visualization as a guide, rather than as a direct producer of quantitative results, *ProkaryMetrics* can be applied to input data across a wide range of imaging modalities, resolutions, histological approaches, and ultimately absolute quality, without requiring modification of the algorithm, or adaptation of numerous parameters. By guiding the user to make specific quantitative measures, rather than relying on subjective expert assessments, *ProkaryMetrics* can be applied by users with widely varying expertise levels, and still produce quantitatively comparable results.

In this manuscript we validate *ProkaryMetrics* for quantitating two pathogen state variables of known clinical importance. The first is the morphology of the organism. The ability of the microorganism to alter its size, by regulation of cell division, provides advantages during disease. "Morphological plasticity" is a well-known survival strategy for fungal pathogens. Its utility for bacterial pathogens is becoming evident through studies of persistence of uropathogenic *Escherichia coli* [6] (UPEC), and Mycobacterium tuberculosis [4]. In addition to being resistant to the host immune response, filamentous morphotypes of organisms are typically resistant to antibiotics even when their non-filamentous progeny are sensitive. This inherent resistance to killing underscores the importance of determining the prevalence of the filamentous morphotypes in infected samples, as successful treatment regimens must be adapted to eliminate these tenacious survivors. The second is the morphology and architecture (in terms of coherent organization) of the pathogenic community. While only gaining widespread acceptance in the past decade, pathogen community architecture, and internal and external organization, have become well understood as directly modulating the effect and effectiveness of treatments for

[*] e-mail: dabdoub.2@osu.edu
[†] e-mail:ray.29@osu.edu

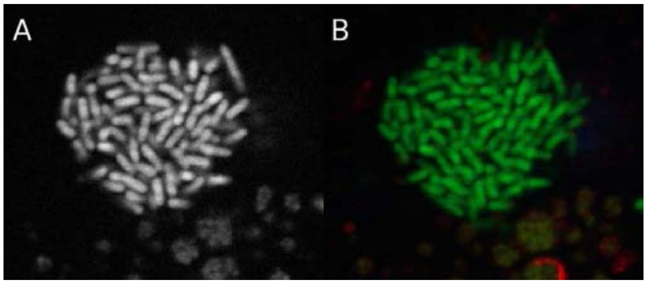


Figure 1: Optical sections of UPEC infected murine bladder biopsy. (A) The green channel of a UTI89 wild type IBC. (B) The same IBC reimaged with a lower-resolution, 3-color modality.

infection.

By leading the user to recognize the general structure and organization of a pathogenic community – a task that can be accomplished with a generalized visualization algorithm that is adequate for descriptive rather than quantitative purposes – and enabling the user to then make specific quantifiable assessments and measures within this generalized presentation, *ProkaryMetrics* provides a mechanism for quantitation of these population-state traits that can be consistently applied with existing technology in typical clinical settings. Our results demonstrate that this approach is adequate to differentiate uropathogenic *E. coli* morphotypes and strains, and to extract quantitatively comparable assessments of community state variables, using both state-of-the-art research, and typical clinical microscopy samples.

2 Software Design

The guiding requirement for *ProkaryMetrics* has been, and continues to be, that the system enable end-user practitioners, without programming experience, and with varying imaging modalities at their disposal, to make quantitatively similar assessments of pathogen/community state traits. As such, the system presupposes nothing more than that the user can acquire one or more digital images of the infected tissue, at sufficient resolution and quality that the user can differentiate individual members of the community by visual inspection. Simple volume segmentation algorithms, and volume visualization approaches enable the user to interactively explore and annotate the pathogen population within the images, with as much dimensional detail as is available in the images themselves. These algorithms and approaches are required only to provide guidance and support for the user's identification of pathogen features, and to quantitatively report the identifications. They are at no point required to automatically determine quantitative features without user guidance.

While this assisted-manual-analysis methodology requires user interaction for every diagnostic analysis, we propose that this is entirely appropriate for clinical applications. Not only does this facilitate timely analysis of samples that are not amenable to automated approaches, it is a practical requirement that any automated clinical diagnostic based on data with quality as variable as microscopy imaging, must be confirmed by inspection by a human expert. Since such inspection is necessarily visual, it is no impediment that our *ProkaryMetrics* approach starts with this process.

3 Software Implementation

ProkaryMetrics is written entirely in Python, relying on the Visualization Toolkit (VTK) [8] to enable visualization of and interaction with volumetric data. Users begin by loading volumetric image data, typically as a series of single-channel 2-dimensional image slices of the sample to be studied. This data is first smoothed using a Gaussian filter, and then isosurfaced with a user-modifiable target pixel intensity value. This volumetric surface rendering is displayed in the visualization window allowing the user to manipulate and explore the data in 3D. The system provides a cursor controlled by the mouse that attaches itself to the nearest rendered surface through ray tracing. In this manner, the user can simply click on a surface outlining a bacterium to place a spherical marker object. One or more markers can then be recorded as representing a bacterium.

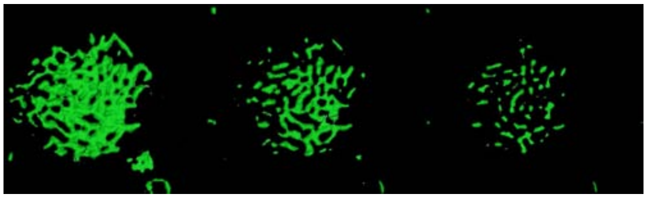


Figure 2: No single isosurface value is appropriate for automated analysis. However, a human assisted analysis of a range of isosurfaces, as shown here, produces nearly identical results for both data sets.

4 Analysis

In order to allow comparison between samples, *ProkaryMetrics* supplies a suite of numeric and statistical tools for investigating the mathematical properties of the biofilms. Using the Khachiyan method [7] for calculating a bounding ellipsoid $E \subseteq \mathbb{R}^n$ for a set of m points, users are enabled to estimate the volume of space occupied by the biofilm under investigation. Furthermore, assuming a standard width and depth for the bacteria, and the length provided by the user, we can estimate the total volume occupied by the sum of the individuals. Combined with the ratio of the two volumes we can present a quantitative picture of the size, shape, and relative packing density of the bacterial community (see Figure 3).

Using the main axis vector of each bacterium, we calculate its scalar projection in the direction of each of the orthonormal basis vectors in $\mathbb{R}^3$. Gathering this information for all of the recorded bacteria, we can compile statistics on the overall layout and orientations of the bacteria within a community. We use these three projections to set the RGB components of the diffuse color of the corresponding bacterium. The resulting visual representation (Figure 4) provides a clear indication of the general orientation trends of the community organization, as well as a means to compare visually between different samples.

Finally, using the midpoint of each bacterium, we borrow a technique from the field of data mining to calculate the average interpoint Euclidean distance (Equation 1), giving another measure of bacterial community packing and another means to compare between communities.

$$d(x_1,x_2) = \sqrt{\sum_{j=1}^{m}(x_{ij} - x_{2j})^2} \qquad (1)$$

5 Results

We have applied *ProkaryMetrics* to the visualization and analysis of UPEC, the major causative agent of urinary tract infections (UTIs). UPEC causes both acute and recurring (mainly in women) UTIs, and results in billions of dollars in medical costs and lost productivity annually [5]. These infections are particularly difficult to treat because UPEC has evolved highly effective means for evading host defenses, as well as medical treatment. The major components of their evasion strategy centers on intracellular invasion of the superficial epithelial cells of the host bladder and morphological change by filamentation. During filamentation, bacteria continue to

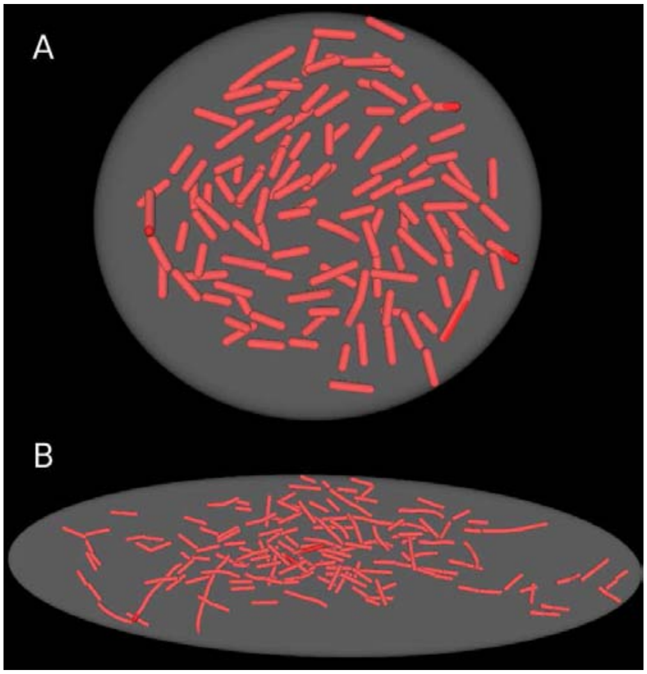


Figure 3: Minimum volume bounding ellipsoids calculated for the user-marked bacteria in (A) a wild type UTI89 IBC and (B) a capsule mutant (UTI89$\Delta kpsF$) IBC. The wild type fills a volume of space approximately $1.95x10^3 \mu m^3$ and is nearly perfectly circular in the x and y dimensions, making it an oblate spheroid. The kpsF mutant is much larger, occupying a volume of $3.87x10^4 \mu m^3$, with a clear bias in one dimension.

grow but are unable to complete the process of division, producing a long strand of conjoined bacteria up to approximately $70 \mu m$ [5]. Establishment of intracellular bacterial communities (IBCs) allows UPEC to avoid the hostile environment of the bladder until the late stages of infection when the host cells begin to porate and apoptose[5]. At this point, the bacteria are exposed and filaments are resistant to neutrophil and macrophage killing, as well as antibiotic treatment [5]. Additionally, the organism traverses distinct stages of development during the infection cycle, three of which are specific to biofilm establishment and growth, and during which the changes in morphology occur [5].

In order to highlight the capabilities of *ProkaryMetrics*, we focus on two UPEC data sets, each containing a single IBC and at least 100 individual bacteria. The first data set is of the wild type UTI89 (a clinical isolate), and the second is a mutant of UTI89 with a defect in the production of capsular polysaccharides, specifically the *kpsF* gene. This capsule mutant is known to produce visually distinct IBCs in size, shape, cohesion, and apparent early onset of morphological change. As with other mutants, viewers could visually comprehend the differences as compared to the wild type, but were limited to vague qualitative descriptions.

In Figure 3 we have used *ProkaryMetrics* to estimate the volume of space occupied by the mass of each IBC, wild type and capsule mutant. By visual inspection, the $\Delta kpsF$ mutant IBC is clearly much larger and less regular in diameter. Fitting an ellipsoid to the data, as is seen in the Figure, we calculate the wild type fills a volume of $1.95x10^3 \mu m^3$ and, within tolerance, fits the category of an oblate spheroid (two of the radii are equal). The $\Delta kpsF$ mutant fits with the visual inspection and has occupies a much larger volume of space of $3.87x10^4 \mu m^3$.

The question of orientation is an interesting one, and certainly important when regarding biofilms. In fact, it is their structure as a community and the spatial heterogeneity of the individuals that contributes greatly to their role as a common cause of persistent infection and their ability to resist treatment [2]. However, obtaining such important information is impossible with destructive techniques such as PCR. As we described in Section 4, the orientation of the main axis vector running along the length of each bacterium is compared to the three orthonormal basis vectors in $\mathbb{R}^3$. In Figure 4, we have used the three orientation calculations to fill the RGB components of the color for each bacterium. *ProkaryMetrics* currently provides three different, user selectable, coloring schemes, and Figure 4 displays the orientation to color mapping: x→blue, y→green, z→red (only bacilli are colored by orientation). In both data sets the bacteria are nearly perfectly aligned with the plane of the image, indicated by the general lack of red in the bacilli. Additionally, with this visualization, it is immediately obvious that the $\Delta kpsF$ mutant is largely dominated by orientation relative to the y-axis (Figure 4B is rotated 90°).

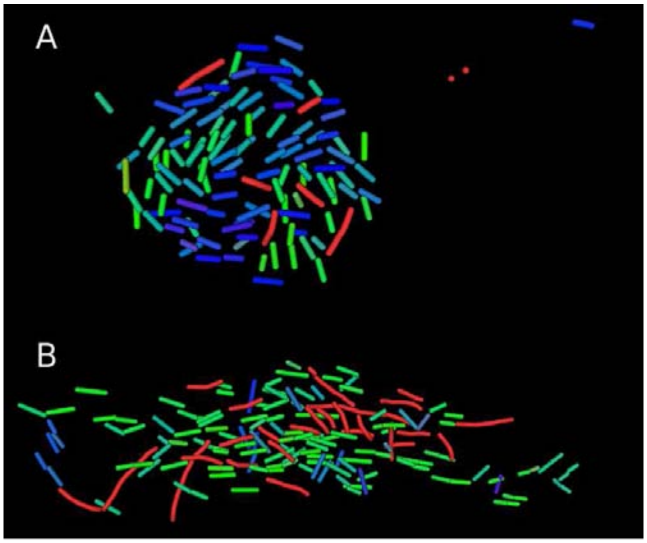


Figure 4: Visual representation of the orientation of the main axis of each bacterium with respect to the three orthonormal basis vectors in $\mathbb{R}^3$. The xyz components of the orientation are represented in the RGB channels of the image: x→blue, y→green, z→red. Filaments and coccoid bacteria are left in their original color. (A) The UTI89 w.t. IBC is clearly split into two fairly distinct populations with one having the main component of its orientation in the y (green) axis, the other with the main component in the x (blue) axis. (B) The IBC formed by the $\Delta kpsF$ mutant is dominated by bacteria oriented along the y axis, indicated by the predominance of green. The image is rotated such that the y-axis is presented horizontally.

While Figures 3 and 4 provide important semi-quantitative analysis regarding overall biofilm organization and architecture, mathematical and statistical comparisons of architectural characteristics are necessary to establish quantitative descriptors that can be used to prove key differences. Table 1 gives a summary comparison of such data, as well as additional information that is not provided by the previous visual representations. Case-in-point, Figure 5 establishes statistical proof of the obvious qualitative differences visible in the orientation visualization in Figure 4. As we would expect, the bacteria in both samples exhibit almost no orientation change in the plane of the image (z-axis). However, both the x-axis and y-axis orientation data show significant differences, with the wild type strain exhibiting similar orientation distribution in both, while the $\Delta kpsF$ strain is dominated by orientation difference in the y-axis (as we expected from Figure 4B by the majority of green-hued bacteria).

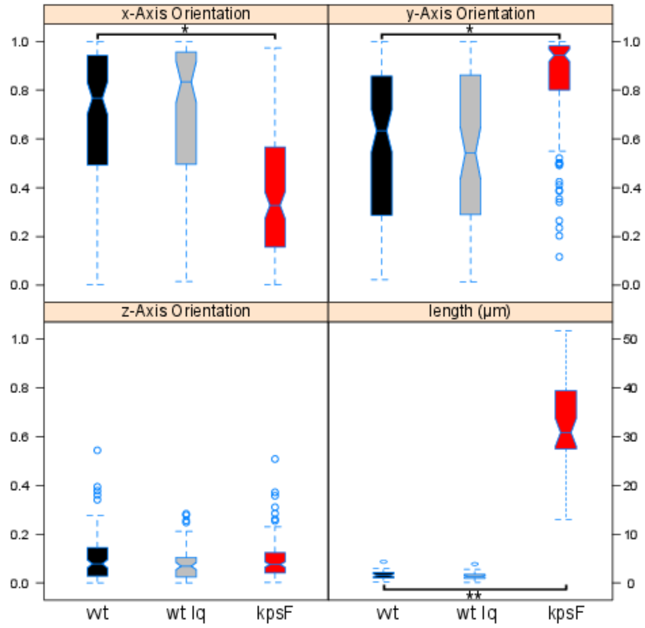


Figure 5: Quantitative representation of the orientations and lengths of the bacteria within the w.t. and $\Delta kpsF$ UTI89 IBCs. The relative orientations are calculated as the projection of each dimensional component on its respective axis. In both, as noted in Figure 4, the relative orientation of the bacteria with respect to the z-axis is nearly flat. Both the x and y-axis projections, however, show significant differences, as do the overall bacterial lengths between the two samples. (*) 2-tailed Mann-Whitney U test, $p < 0.0001$ (**) unpaired 2-tailed Student's t-test, $p < 0.0001$. Finally, the gray barplots in each graph represent the data gathered from the lower quality data set seen in Figures 1B and 2. Despite the loss of information, we were able to achieve nearly identical quantitative results with *ProkaryMetrics*.

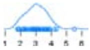
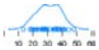
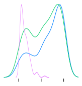
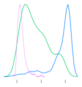
Quantifier	UTI89 *w.t.*	UTI89$\Delta kpsF$
Count	120	161
Length (μm)		
IBC Volume (μm^3)	1949	38660
IBC Diameter (μm)	21.6 19.6 3.31	32.0 93.8 12.9
Volume Ratio	0.58	0.32
Orientation (0.0-1.0)		
IB Distance (μm):	15.7	62.5

Table 1: This table summarizes the quantitative descriptors we have developed for architectural comparisons of biofilm infections. The volume ratio is the total volume of bacteria to the IBC volume. The orientation histograms are colored green, blue, pink for the x,y,z orientations respectively.

6 Conclusion

It is becoming clear that in many cases, current methods for clinical investigation of biofilm infections are either insufficient or inaccurate due to the amount of time required or their destructive nature. An integral aspect of the nature of biofilms is their overall architecture as well as the arrangement of the individual pathogens. Indeed, in UPEC distinct architectural and morphological changes occur through the three stages of its intracellular growth cycle [5]. Understanding and calculating these various properties in a quantitative manner can be important for identification of disease state and potential susceptibility of an infectious organism. A non destructive 3D microscopy-based tool is ideal for meeting these needs. Here we present new software, *ProkaryMetrics*, as a tool to fit these requirements, providing 3D visualization and qualitative, as well as quantitative, analyses for user-assisted identification of bacteria from volumetric microscopy data.

In order to demonstrate the utility of the software, we have applied *ProkaryMetrics* to the visualization and analysis of a model organism that is recognized as the causative agent of most urinary tract infections: UPEC. As an infectious organism, it prefers intracellular existence during the majority of its lifecycle. While intracellular, it forms biofilm-like structures (IBCs) that are necessary for pathogenesis [5]. Applying the software to an IBC of wild type UTI89 and the UTI89$\Delta kpsF$ mutant, we have established significant qualitative and quantitative differences between them in overall architecture and individual characteristics including IBC volume and shape, as well as aggregate and specific orientation and length parameters. While the software was developed with UPEC in mind, it is generalizable and easily modified to handle any sort of organism. Currently, in addition to UPEC analysis, we are generating analyses in collaboration with researchers studying non-typeable *Haemophilus influenzae*. Furthermore, we are investigating computational image analysis techniques for partially automated processing of microscopy data to aid the user in identification [3].

Developing such algorithms, descriptors, and objective analyses is necessary for accurate identification and comparison of clinical biofilm samples in all stages of infection. Such non-destructive imaging analyses will provide rapid and important guidance for clinicians and improve the suite of tools available for disease assessment.

References

[1] J. W. Costerton, J. C. Post, G. D. Ehrlich, F. Z. Hu, R. Kreft, L. Nistico, S. Kathju, P. Stoodley, L. HallStoodley, G. Maale, G. James, N. Sotereanos, and P. DeMeo. New methods for the detection of orthopedic and other biofilm infections. *FEMS Immunology & Medical Microbiology*, 61(2):133–140, Mar. 2011.

[2] J. W. Costerton, P. S. Stewart, and E. P. Greenberg. Bacterial biofilms: A common cause of persistent infections. *Science*, 284(5418):1318–1322, 1999.

[3] S. M. Dabdoub, S. S. Justice, and W. C. Ray. A dynamically masked gaussian can efficiently approximate a distance calculation for image segmentation. *Advances in Experimental Medicine and Biology*, 696:425–432, 2011.

[4] K. England, R. Crew, and R. Slayden. Mycobacterium tuberculosis septum site determining protein, ssd encoded by rv3660c, promotes filamentation and elicits an alternative metabolic and dormancy stress response. *BMC Microbiology*, 11(1):79, 2011.

[5] D. A. Hunstad and S. S. Justice. Intracellular lifestyles and immune evasion strategies of uropathogenic escherichia coli. *Annu Rev Microbiol*, 64:203–21, Oct 2010.

[6] S. S. Justice, D. A. Hunstad, L. Cegelski, and S. J. Hultgren. Morphological plasticity as a bacterial survival strategy. *Nat Rev Micro*, 6(2):162–168, Feb. 2008.

[7] L. G. Khachiyan. Rounding of polytopes in the real number model of computation. *Mathematics of Operations Research*, 21(2):307–320, May 1996.

[8] W. Schroeder, K. Martin, and B. Lorensen. *Visualization Toolkit: An Object-Oriented Approach to 3D Graphics, 4th Edition*. Kitware, 4th edition, Dec. 2006.

Trauma Analysis through Data-Driven Medical Injury Visualization

Patrick J. Gillich

U.S. Army Research Laboratory

Abstract

Scientists of all disciplines work in both the spatial and non-spatial realm, and require visualization of data early in the process of discovery. Visualization of multi-dimensional human trauma data greatly enhances the communication and examination of data for analysis. Translating medical coding into pictures enables analysts to examine visual data to spot patterns, trends, outliers, and to generally gain an increased understanding. A graphical tool named the Visual Anatomical Injury Descriptor (Visual AID) enables individuals to illustrate injury onto an anatomical figure and perform discovery operations by inspecting injury patterns using composite information.

Keywords: Injury analysis, abbreviated injury scale, injury scoring, wounds, trauma, and injury

Index Terms: I.3.2 [Computer Graphics]: Graphics Systems— Stand-alone systems; K.8.1 [Personal Computing]: Application Packages — Graphics

1 Introduction

Every branch of science needs to observe its unique phenomena and each has its own specialized techniques for measuring and collecting representative data. Many observed phenomena have a meaningful, intrinsic spatial component. The spatial components are often coupled to greater amounts of non-spatial components for information discovery. To fully observe these, one needs the subject matter expertise and instruments to measure and collect data as well as the tools to visualize them.

Visualizations often produce appealing images that attract readers to accompanying text in proximity. However, in scientific disciplines that work in both the spatial and non-spatial realm, visualization of data is useful very early in the process of discovery. Translating numbers into pictures enables analysts to examine visual data to spot patterns, trends, outliers, and to generally gain an increased understanding. Data and analysis are communicated more efficiently and effectively to a broader audience through the use of illustration.

Visualization of multi-dimensional human injury data greatly enhances the communication and examination of trauma data for analysis. A graphical tool that allows individuals to illustrate injury onto an anatomical figure has been constructed to support this need. This tool is the Visual Anatomical Injury Descriptor (Visual AID).

Street Address and Electronic Mail Address

LEAVE 0.5 INCH SPACE AT BOTTOM OF LEFT COLUMN ON FIRST PAGE FOR COPYRIGHT BLOCK

2 Overview

Visual AID is a computer graphics tool developed by the U.S. Army Research Laboratory to improve injury visualization and effectively communicate trauma described by medical or simulated data. Data visualization in Visual AID uses novel techniques to represent the authentic relationships of source injury data. Injury visualization is meant to communicate particular insight and knowledge that is evident in the underlying data. There are two main purposes for using Visual AID. The first is to create visuals that illustrate known injuries for the purposes of presentation. The second is to create visuals to aid in analysis and data examination of the previously unknown. This allows the analyst to facilitate discovery and identify trends or relationships of note.

2.1 Background

Visual AID was developed following the success of several three-dimensional human trauma images that were created by hand. These images received acclaim from the community and confirmed the value of high-resolution anatomical depictions. The success of these images was due to their visual portrayal in the spatial domain and the ease with which they could be understood. Previously, images used for injury description consisted of the skeletal structure with a body outline, on which injuries were manually marked with red dots.

The success of the initial illustrations prompted an immediate demand to provide such an improved visualization capability to a wider audience. Generation of the new illustrations required three-dimensional modeling expertise and significant time for construction. This process and the complexity and cost of the graphics development environment were a significant technical burden to the end users. A general-purpose tool that uses CAD geometry in a controlled and limited manner was the preferred option.

2.2 Application

Visual AID is a powerful, adaptable, and user-friendly application that enables the efficient and effective communication of trauma data. Its capabilities are driven by user input described either by named anatomical locations or by a standard anatomical-based injury classification system. The user can enter collections of patient records and the tool will perform frequency analysis across body regions and types of anatomical structures. Visual AID creates illustrations in real-time onto a reference anatomical figure, providing a quick and easy technique for describing trauma data for analysis and information reporting.

Visual AID is a specialized tool for the non-3D modeler that replaces a complex and tedious process that required expert knowledge of 3D modeling and visual effects. Analysts with medical records or simulated injury data are using the tool to readily create detailed illustrations. Certified injury scorers are using the tool to aid in coding and in the validation of coded results. Medical and injury scoring educators can use the tool to communicate the significance of trauma and teach injury classification.

3 Injury Classification

Anatomical-based injury classification systems describe the impact of an injury in terms of the extent of tissue damage and generally define severity in terms of threat-to-life. Developed for both patient triage for medical emergencies and as a predictor in evaluating the impact of services or systems on patient survival, these indices are currently being used to characterize survivability of personnel in civilian and military scenarios. Injury type and severity classification is critical for the evaluation of systems that have requirements for personnel protection or mitigation of injury.

Visual AID integrates the current version of the Abbreviated Injury Scale (AIS), the standard lexicon for coding individual anatomical injuries that includes a consensus-derived estimate of severity [1]. The AIS is the most widely used anatomical-based injury classification system for characterizing personnel trauma. This scoring system was introduced in 1971 and its current version is AIS-2005 update 2008. This version contains approximately 2000 injury descriptors. The AIS was designed to distinguish between types of trauma of clinical importance as well as types of trauma of interest to system designers and research engineers. It is a valuable tool in the scientific study of the epidemiology of trauma and trauma outcomes. It is used by the government, academia, and industry for vehicle blast test evaluation, vehicle crash investigation, clinical trauma research, and is applied directly to records in trauma registries.

AIS-2005 is the culmination of the collaborative efforts of many individuals from different disciplines, organizations and nations. It is a major update of AIS that expands the number and sophistication of injury descriptors, uses more modern nomenclature, and captures subtle variations in injury. The AIS classifies injuries across all body regions and types using a time independent code. It considers only the injury and not its consequences, with a few exceptions that include blood loss and loss-of-consciousness. Hence, time dependent complications, such as infection, are not classified.

AIS-2005 codes consist of a six-digit injury descriptor, which is unique for each injury description, followed by a single-digit severity score. A severity score is assigned to each injury descriptor, using a six-point ordinal scale with levels that range from 1 (relatively minor) to 6 (maximal or virtually unsurvivable) (see Figure 1). This severity score is based on, but not limited to, several components: threat-to-life, hospitalization requirement, treatment complexity, treatment cost, treatment duration, permanent impairment, and quality-of-life. The scale of the severity score is not linear, in the sense that the difference between a severity of 1 and 2, is not equivalent to the difference between a 3 and 4. Therefore, it is improper to average AIS severities.

A full AIS code for a given injury has seven numerals. For example, the code 440606.3 represents a solitary diaphragm laceration less then or equal to 10 cm in length. The first 4 indicates thorax as the body region, the 4 in the second position indicates an organ; the next two digits are context specific to the first two, where the 06 in places three and four indicate diaphragm; the next two digits are context-specific to the first four, where the 06 in the fifth and sixth positions, indicates a laceration less then or equal to 10 cm in length. The 3 in the final position is the severity score and specifies a serious injury.

4 Capabilities

Visual AID is used to perform injury analysis based on information from patient medical records or as a result of modeling and simulation. Several different types of injury analysis can be performed, including tissue damage analysis, which uses injury data with limited resolution, injury severity analysis, which uses higher resolution information to examine single or multiple casualties, and injury frequency analysis, which examines injury relationships in multiple casualties for a collection of trauma-causing events.

AIS Severity	Injury Level
1	Minor
2	Moderate
3	Serious
4	Severe
5	Critical
6	Maximal

Figure 1. AIS Severity Scale

4.1 Tissue Damage Analysis

Injury type and severity classification is a time-consuming effort that requires specialized, trained personnel. In support of limited or preliminary injury descriptions, Visual AID can be used to visually identify damage to specific anatomical structures without specifying the nature of the injury and severity. It uses the color blue as a single severity-independent color to describe damage on visual anatomical depictions. This allows trauma data to be reported rapidly to data consumers in the spatial realm while using gross descriptions of the non-spatial elements. These types of images are typically used when patient injury details contain only the specific anatomical structures damaged and lack the nature of the injury and severity. This occurs when the necessary details are not available or have not been finalized. An example use case is when autopsy cases have not been finalized and only preliminary information is available.

The images in Figure 2 are examples of tissue-damage injury illustrations. Cross-hatching is used to indicate complete tissue loss. Highlighting of body structures is used to indicate an injury to that structure.

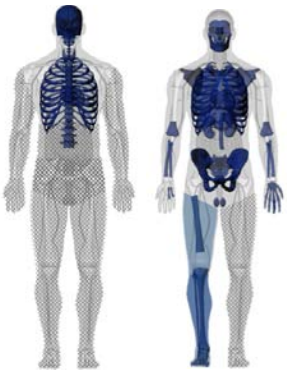


Figure 2. Tissue Damage Analysis Illustrations

4.2 Injury Severity Analysis

Given injury type and severity classification, Visual AID can be used to visually illustrate damage to specific anatomical structures. Visual AID creates visual anatomical depictions using a color palette representing the six AIS severity levels shown in Figure 1. This allows trauma data to be reported in the spatial and nonspatial realms, where elements of anatomical structures are highlighted to indicate damage. This use of visualization encodes information for a given injury and represents it through placement, color, label and size. The image in Figure 3 is an example of an injury description.

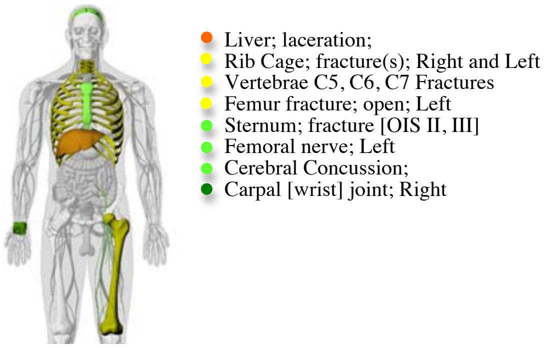

● Liver; laceration;
● Rib Cage; fracture(s); Right and Left
● Vertebrae C5, C6, C7 Fractures
● Femur fracture; open; Left
● Sternum; fracture [OIS II, III]
● Femoral nerve; Left
● Cerebral Concussion;
● Carpal [wrist] joint; Right

Figure 3. Injury Severity Analysis Illustration

4.3 Side-by-Side Patient Analysis

For trauma-causing events involving multiple patients, Visual AID can perform event-level visualizations to examine injury relationships. A side-by-side visual comparison, as illustrated in Figure 4, shows patient cases associated with a single event. These cases can be grouped together by user-defined criteria. Each patient is represented by a scaled-down image to support the display of multiple figures adjacent to each other on the screen. Injury commonality can be easily examined to identify similar patterns by body region (i.e., thorax), type of anatomical structure (i.e., skeletal), specific type of anatomical structure (i.e., sternum) and severity (i.e., moderate). The benefit of this technique is that it enables an event to be summarized pictorially, while at the same time allowing the user to interact with the underlying data to better examine event-level patterns.

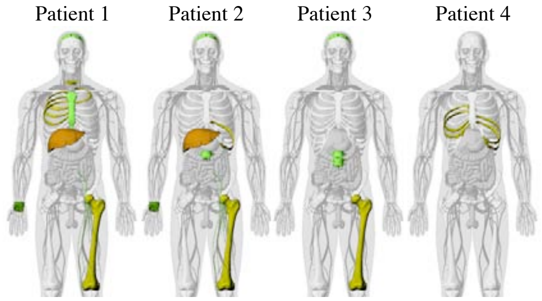


Figure 4. Side-by-Side Patient Illustrations

4.4 Cumulative Frequency Analysis

In addition to comparing sustained injuries side-by-side, sometimes it is useful to categorize and count injuries over a series of patients or events. Visual AID includes this capability in the form of a frequency analysis mode, as illustrated in Figure 5. Injury frequency can be examined across the total number of patients or injuries. In this mode of operation, a unique frequency color palette is used to illustrate data density. Injuries are categorized by nature-of-injury and body region (e.g., traumatic brain injury, spinal cord injury, vertebral column injury, torso and extremities).

Frequency of injury is illustrated either by body region or type of anatomical structure. The user can control what is visualized through filters on severity and type of anatomical structure, and perform data queries to keep and remove aspects of the patient data records.

Standardized medical data selection and reporting combined with several unique visualization techniques allows Visual AID to classify injury by type and anatomical region in a novel manner. This functionality provides a manageable number of clinically meaningful diagnostic categories that characterize nature-of-injury and body region categories. This process: a) simplifies the process of classifying injuries; b) provides a standard format for reports; c) serves as a standard for comparative studies; d) characterizes patterns of injury.

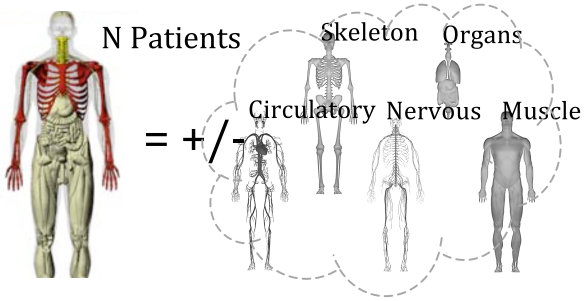


Figure 5. Frequency Analysis Illustration

5 Conclusion

The Visual AID tool has been implemented and is currently being used to visualize and analyze injury, with associated threat-to-life, and the distribution of wounds and underling injuries. It uses AIS-2005 update 2008, a precise anatomical injury classification standard, as its foundation of anatomical injury scoring. Visual AID allows visualization of wounds and associated injuries onto a three-dimensional anatomical model of the human body and allows the user community to easily create detailed injury illustrations.

References

[1] International Injury Scaling Committee. *Abbreviated Injury Scale (2005 Revision Update 2008)*. Arlington Heights, IL: American Association for Automotive Medicine, 2008.

TAO: Terrain Analytic Operators for Expert-Guided Data Mining Applications

Jason McLaughlin[1] Qian You[1,2] Shiaofen Fang[1,2] Jake Y. Chen[1,2,3,4]

[1]Department of Computer and Information Science, Purdue School of Science, Indianapolis, IN 46202
[2] Center of Bio-computing, Indiana University-Purdue University, Indianapolis, IN 46202
[3] Indiana University School of Informatics, Indianapolis, IN 46202
[4] Indiana Center of Systems Biology and Personalized Medicine, Indianapolis, IN 46202

ABSTRACT

We present an iterative visual analytic method for exploring large multivariate datasets in biological applications. By integrating terrain visualization, an interactive interface, and an iterative refinement model, a user can integrate their domain knowledge to explore large datasets based on individual analysis objectives. This domain neutral visual analytic system facilitates a reason-based understanding of large, abstract datasets. A case study in biomedical gene expression analysis demonstrates effective and interpretable data mining results.

KEYWORDS: Visual analytics; machine learning; information visualization

INDEX TERMS: H.3.3 [INFORMATION STORAGE AND RETRIEVAL] Information Search and Retrieval – Query formulation and relevance feedback; H.5.2 [INFORMATION INTERFACES AND PRESENTATION] User Interfaces – user-centered design;

1 INTRODUCTION

A primary reason for processing data is to discover hidden knowledge for better decision making or problem solving, but this rationale is often decoupled from computational analysis. Human beings and computers have complementary advantages in information processing that can be integrated using Visual Analytics (VA), a developing "science of analytical reasoning facilitated by interactive visual interfaces" [1].

VA is useful when massive amounts of data not only overwhelm the analysts, but also when traditional data analysis and mining techniques fall short. Automatic data analysis or mining models search for optimal solutions after the objectives of the computing tasks are defined. However, for many of today's data sets, the meaningful patterns and hidden knowledge are not known beforehand, hence it is hard to formulate the goals of discovery in the first place. VA can leverage human perception, intelligence and reasoning capability, and cooperate with high-speed computing to solve complex problems.

We present several Visual Analytic operators that allow a user to interactively explore related data that are responsive to different input features. A novel 3D terrain visualization provides an intuitive overview of the effects of different features on the most relevant data objects. An iterative refinement model allows the user to gradually change the input features, asses their effects, and continuously improve their learning model by using Terrain Analytic Operators. Since the user is constantly involved in each iteration of the model, their understanding of the data can evolve to clarify their analysis.

We apply our Iterative Visual Analytic system to the study of gene expression patterns in human breast cancer molecular networks. Cancer biomarker panels of differentially expressed genes in molecular interaction networks are analyzed to identify a simpler characteristic signature gene subnetwork group that is sensitive and specific for identifying breast cancer. Such characterization, which would have involved careful design of machine learning methods tailored to the cancer gene expression and cancer molecular network data sets, can help biologists with little formal computer training, yet vast biological domain knowledge, generate high-quality hypotheses quickly.

2 RELATED WORK

Visual Analytics is a relatively new field that seeks to facilitate analytical reasoning through the use of visual user interaction. It emphasizes a tightly coupled interaction between the user and machine to solve problems that may be otherwise intractable [1]. It is a multidisciplinary field and is related to information and scientific visualization and data mining.

It is often desirable to reduce a large set of multivariate features to a more manageable predictive subset. Previous automated methods for finding satisfactory feature subsets in both supervised and unsupervised machine learning require a quantifiable metric for evaluating feature sets. When analyzing complex data sets, it is likely that a user's analysis goals may change as they explore the data, and traditional methods are not designed for that.

There have been many techniques for visualizing high-dimensional data sets [2]-[7], but these designs rarely assist users to track the development of associated insights and knowledge. The interactions with these systems are not designed to allow user feedback to effectively drive the underlying data analysis model. To tightly couple interactive visualization with a user's reasoning process remains an early research topic [8]-[10].

3 METHODS

3.1 Framework: An overview

This Iterative Visual Analytic method allows a domain-expert user with minimal formal computer training to seek correlations between a set of feature vectors and a set of data objects, with several objectives in mind:

Multivariate feature selection – The user selects a set of feature vectors and is presented with a meaningful representation of their connection to the current set of data objects.

Allow user to guide search based on domain knowledge – Sets of feature vectors and data objects should represent objects that are familiar to the user and meaningful in their domain. The user should always be aware of how changes to the feature set affect the visualization and be able to easily interpret those changes based on their goals.

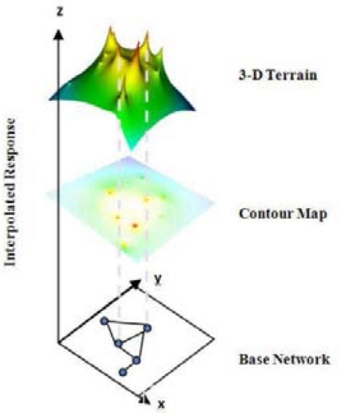


Figure 1 – Terrain Visualization, a 2D base network of associated terms and interpolated height values based on each nodes response to an input feature

Clearly visualize relevant correlations - A user should be able to easily identify which data objects correlate strongly with the current feature vector. It should be clear whether the correlation is specific to individual data objects or represents a response from many neighboring data objects.

Provide domain-neutral tools for use in domain-specific applications - A 3D Terrain visualization represents a set of associated domain-specific terms and their response to a set of features. Using TAOs, the user constructs and compares terrain visualizations to modify and analyze their input features. This robust visualization and analysis technique allows the user to assess the effects of different feature sets on a wide variety of terms of interest. Feature sets are iteratively selected and refined by the user to improve the sensitivity and specificity for terms of interest.

3.2 Terrain Visualization Techniques

Our 3D visual terrain represents the response of a network of associated domain terms to an input set of features.

1. Networks of domain terms are represented as a node link graph and arranged on a 2D layout. Nodes represent a related set of terms that are dependent on the domain application, and weighted edges represent any hypothesized associations between those terms. The graph layout algorithm uses a node-weight edge-weight method of GeneTerrain [12]

2. For a single feature vector, we calculate a scalar response variable for each node. For every point on a regular grid on the 2D layout, we use Shepard's interpolation to calculate a height value based on the response variable of surrounding nodes: Using VTK, we render a terrain based using those height values.

For multiple feature vectors we calculate a consensus terrain, where the height of each point in the response terrain is the mean average of the heights for each of the single-feature vector terrains.

3.3 Terrain Analytic Operators for Iterative Visual Refinement

The Terrain Visualization described above allows a user to quickly assess the relatively predictive power of a set of features for a group of related nodes. Our Iterative Visual Refinement Model allows the user to search for a set of features with improved prediction rates.

The model iteratively adds or removes features from the current candidate group. It uses 3 TAOs to loop through three steps: TAO_CONSTRUCT, TAO_COMPARE, TAO_MODIFY (TAO_ADD and TAO_REMOVE). The biologist can stop the search once an effective set of features is found.

1.TAO_CONSTRUCT - Update and Visualize Terrain:

Input – Set of one or more feature vectors

Function – Construct terrain as described in 3.2. The term association network is the base network and the feature is the response variable. If there is more than one feature in the set, construct the consensus terrain of the feature set.

Output – Render and display the terrain.

2. TAO_COMPARE - Assess Performance:

Input – Set of two or more displayed terrains

Function – Visually evaluate the specificity of a targeted node by observing the terrain profile and comparing it to previous iterations. This decision making process is partially subjective and can be based on criteria such as: (a) Whether the target node has a dominant and characteristic landmark features on the terrain. (b) Whether profile noise caused by other nodes is tolerable; although a node may reduce the specificity of other targeted node(s) this noise can be acceptable. (c) The current number of the features in the feature set.

Many more criteria can be added, depending on the user's domain knowledge and specific requirements of the application. The iterative visual analytical process helps the user make flexible decisions and search for an improved solution. At the end of this step, the user will either proceed to step 3 or terminate with current or previous sets.

Output – Set of features corresponding to user-selected terrain.

3. TAO_MODIFY – Alter input features:

Input – Set of features from TAO_COMPARE and a subset of features to be added or removed

Function – Apply TAO_REMOVE or TAO_ADD to current feature set

Output – Modified set of features.

TAO_REMOVE - A series of newer signatures is obtained by removing one feature at one time. For example, given a signature with features A, B, C and D, evaluate signatures of BCD, ACD, ABC.

TAO_ADD - If the user knows that certain features might improve specificity, these can be added for assessment. For

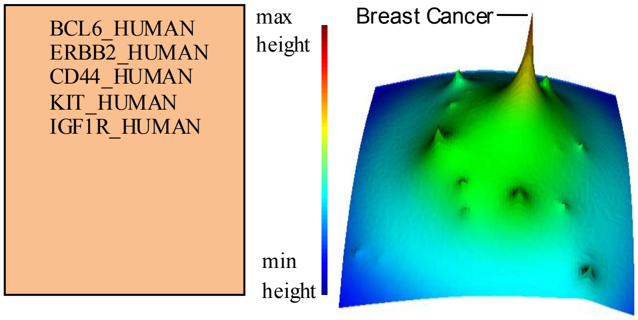


Figure 2 – Terrain Visualization of initial biomarker signature. Diseases are laid out in the x-y plane, where near proximity to neighboring diseases represents higher co-occurrencew ith respect to potential biomarkers. The height of each peak represents a disease's response to the current five-biomarker panel. Color is mapped to terrain height for clearer viewing.

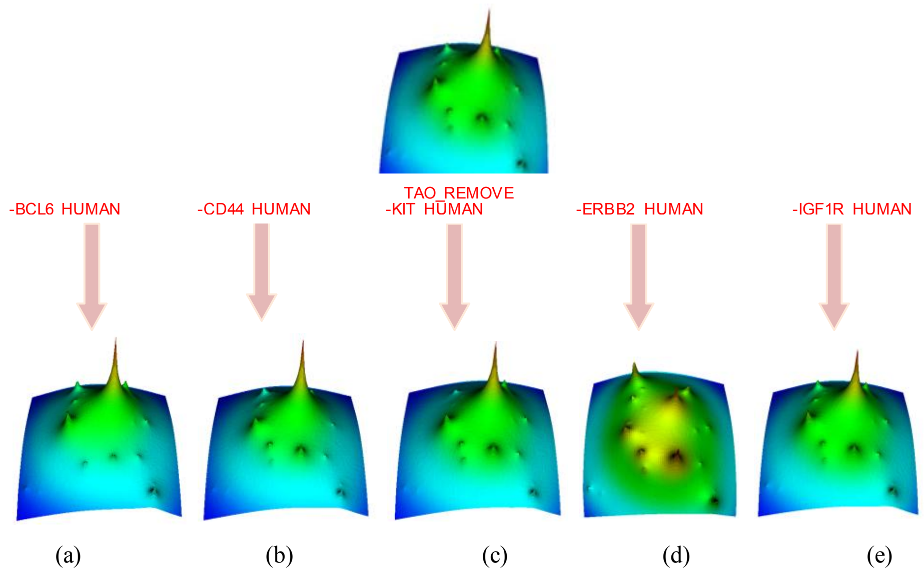


Figure 3 – TAO_REMOVE generates 4-marker panels. TAO_CONSTRUCT renders five terrains, and all results are displayed to the user with TAO_COMPARE.

example, if we suspect that E is a good feature, newer signatures EBCD, EACD, or EABC are given as inputs to Step 1.

Continue iterating from Step 1-3 until satisfied with the results.

4. **Termination**: Render the final terrain of the feature set with the optimal response. Assess the improved classification rates of the new feature set.

4 Biomolecular and Disease Network Terrains

A case study applies our Iterative Visual Analytic framework to biomarker discovery. Networks of diseases are often correlated with gene expression values and a sensitive and specific panel of biomarkers for disease(s) is desirable.

In this case study, to identify a biomarker signature with high disease sensitivity and specificity for breast cancer, we began with a set of 5 biomarkers with the highest individual rank for breast cancer prediction: BCL6_HUMAN, ERBB2_HUMAN, CD44_HUMAN, KIT_HUMAN, IGF1R_HUMAN.We began with cancer association scores for 54 cancer types derived using a method from biomedical literature mining.

In Figure 2, although the profile of the initial feature set has breast cancer as a distinct peak (suggesting high sensitivity), there is a large amount of noise, so the specificity needs to be improved.

Figure 3 shows the resulting disease terrain profiles after the initial signature is modified. Now the user assesses performance. Removing BCL6_HUMAN from the initial signature significantly reduces the noise in the circled region in (a); removing CD44_HUMAN from the initial signature only slightly reduces noise in the circled region in (b); removing KIT_HUMAN from the initial signature slightly reduces noise in the circled region in (c); removing ERBB2_HUMAN from the initial signature greatly introduces more noise over all the profile in (d); removing IGF1R_HUMAN does not result in visible changes in the profile (e).

The above indicates that removing BCL6_HUMAN has the most improved profile due to the significant noise reduction. Therefore we only accept the modification of removing BCL6_HUMAN.

The signature of the current 4 marker panel still does not yield a satisfactory profile, so we continue with the Modify Signature step and constructed terrains for the newer signatures in figure 4 (a)-(d). Removing KIT_ HUMAN (a) results in the best profile. At this point the biomarker panel includes ERBB2_HUMAN, IGF1R_ HUMAN and KIT_ HUMAN. When comparing the disease terrain profiles of the initial signature (5a) and the signature after iterative refinement model (5b), one can see that the newer signature has much less noise in the profile therefore has an improved specificity to the breast cancer.

To further build on the current best signature, two additional signatures are defined by removing IGF1R_HUMAN and CD44_HUMAN. Based on images generated by last two iterations, ERBB2_HUMAN is not removed.

The user observes that Figure 5(b) represents a better improvement than ERBB2_HUMAN and CD4_HUMAN (not displayed). The signature of two biomarkers, ERBB2_HUMAN and IGFIR_HUMAN, is satisfactory and we proceed to Step 4.

The final gene signature ERBB2_HUMAN and IGF1R_HUMAN (Figure 5b) significantly improves the expected disease sensitivity and specificity of breast cancer. The user can now use this biomarker panel for creating a classification system or conducting future experiments.

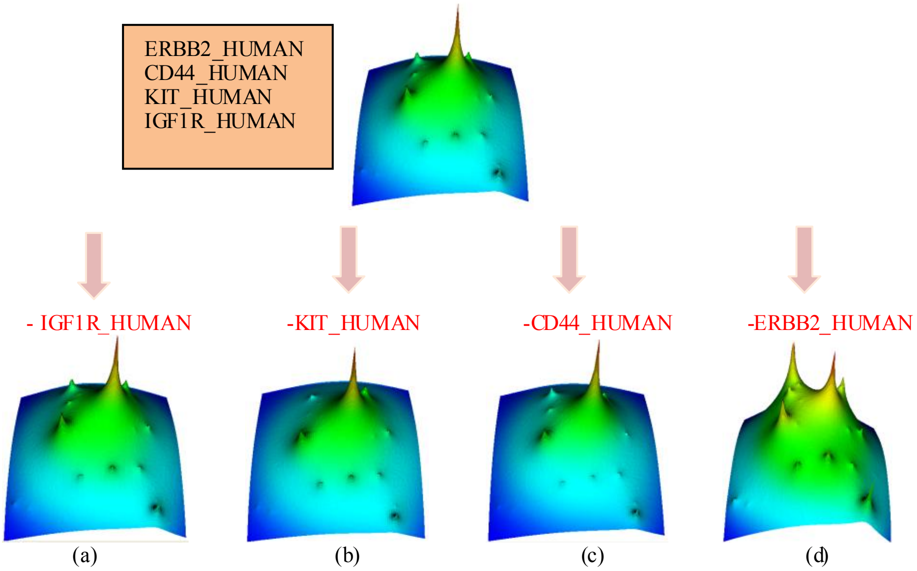
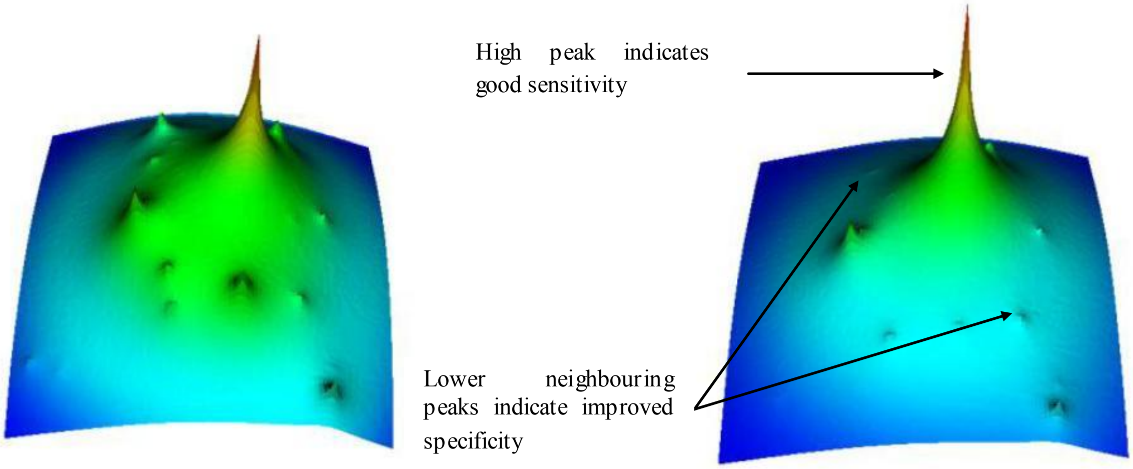


Figure 4 – Additional iteration with 3-marker panels



High peak indicates good sensitivity

Lower neighbouring peaks indicate improved specificity

(a) Initial Disease Terrain from top 5 biomarkers

(b) Final Terrain ERBB2_HUMAN and IGFIR_HUMAN

Figure 5-Improved sensitivity and specificity in final panel

5 Discussion

We introduced a domain-neutral Iterative Visual Analytic framework for large scale data analysis. We demonstrated that 3D terrain visualization can intuitively assess gene biomarkers for detecting Breast Cancer. Terrain visualization has also been shown to be an effective way to visually explore sensitive and specific biomarkers for Alzheimer's disease [12] and correlations among cancer and gene term networks [13]. This interactive system allows a user to apply their domain knowledge to reduce the complexity of a problem and to make reason-based refinements to their analysis as they explore large datasets.

6 Acknowledgements

This project is partly supported by the Open Project Program of the State Key Lab of CAD&CG, Zhejiang University, China.

7 REFERENCES

[1] J. Thomas and K. Cook, Illuminating the Path: The Research and Development Agenda for Visual Analytics, IEEE Computer Society, 2005.

[2] A. Inselberg, "The plane with parallel coordinates," The Visual Computer, vol. 1, pp. 69-91, 1985.

[3] I. Borg and P. Groenen, "Modern multidimensional scaling: Theory and applications," Journal of Educational Measurement, vol. 40, pp. 277-280, 2003.

[4] E. Kandogan, "Visualizing multi-dimensional clusters, trends, and outliers using star coordinates," in Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001, pp. 107-116.

[5] P. Hoffman, G. Grinstein, and D. Pinkney, "Dimensional anchors: A graphic primitive for multidimensional multivariate information visualizations," in Proceedings of Workshop on New Paradigms in Information Visualization and Manipulation, 1999, pp. 9-16.

[6] B. Shneiderman, "Tree visualization with tree-maps: 2-d space-filling approach," ACM Transactions on Graphics, vol. 11, pp. 92-99, 1992.

[7] S. Havre, E. Hetzler, P. Whitney, L. Nowell, B. P. N. Div, and W. A. Richland, "ThemeRiver: Visualizing thematic changes in large documentcollections," IEEE Transactions on Visualization and Computer Graphics, vol. 8, pp. 9-20, 2002.

[8] D. Gotz and M. Zhou, "Characterizing users' visual analytic activity for insight provenance," in Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, 2008, pp. 123-130.

[9] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: Supporting investigative analysis through interactive visualization," Information Visualization, vol. 7, pp. 118-132, 2008.

[10] T. Jankun-Kelly, K. Ma, and M. Gertz, "A model and framework for visualization exploration," IEEE Transactions on Visualization and Computer Graphics, vol. 13, pp. 357-369, 2007.

[11] J. Li, X. Zhu, and J. Chen, "Mining disease-specific molecular association profiles from biomedical literature: A case study," in Proceedings of the ACM Symposium on Applied Computing, 2008, pp. 1287-1291.

[12] You Qian, Shiaofen Fang, and Jake Y. Chen GeneTerrain: Visual Exploration of Differential Gene Expression Profiles Organized in Native Biomolecular Interaction Networks (2008) Information Visualization. March 20, 2010 vol. 9 no. 1 1-12.

[13] Qian You; Shiaofen Fang; Mukhopadhyay, S.; Vaka, H.; Chen, J.; , "Visualizing a Correlative Multi-level Graph of Biology Entity Interactions," Network-Based Information Systems, 2009. NBIS '09. International Conference on , vol., no., pp.304-309, 19-21 Aug. 2009 doi: 10.1109/NBiS.2009.37

The (Inter)Face of Kalm

Halimat Alabi
University of Victoria
3800 Finnerty Road
Victoria, BC V8P 5C2 Canada
1-778-676-0411
halimat@uvic.ca

Dr. David Worling
Westcoast Child Development Group
Suite 203, 3195 Granville Street
Vancouver, BC V6H 3K2
1-604-732-3222
dworling@childdevelopmentgroup.com

Dr. Bruce Gooch
University of Victoria
3800 Finnerty Road
Victoria, BC V8P 5C2 Canada
1-250-483-5941
brucegooch@gmail.com

ABSTRACT

This paper explores the design of Kalm, a mobile anxiety management application. Based on cognitive behavioral therapy (CBT) methodology, the application will be a source of support, education and entertainment for the user. Clinicians may use it to establish a baseline anxiety experience, prescribe and/or monitor CBT homework. The application's interface must present potentially large amounts of data in such a way that it is both intelligible and relevant to the two stakeholder groups, users and clinicians.

We also aim to create an interface that is intuitive and accessible to differently-abled users. This paper explores methodology to analyze, filter and illustrate feedback from the various activities of the Kalm application. The design of the interface for this application will contribute to the design of ubiquitous health applications on mobile devices.

Categories and Subject Descriptors

D.m Miscellaneous Software Psychology, H.m Miscellaneous Information Systems.

Keywords

Personal health informatics, mobile technology, mental health applications, ubiquitous healthcare, persuasive design, gamification, serious games.

INTRODUCTION

At some point everyone has felt the racing heart, sweaty palms or nausea of anxiety – it affects all ages, genders and cultures. When it gains momentum heightening to the level of an anxiety disorder, therapeutic intervention is recommended. Almost 50% of Americans will meet the criteria of an anxiety disorder in their lifetime [2]. Nearly 75% of those with an anxiety disorder have their first episode by age 22, making early intervention crucial [9].

Though empirical studies have found CBT methodology effective, the therapy is often inaccessible to the people who need it due to cost, distance or the availability of a qualified therapist [6]. This results in many people with anxiety disorders – both subclinical and clinical – not receiving appropriate assistance. The burden of illness for anxiety is far reaching; suffers are at increased risk for depression, substance abuse and suicide if the condition is left untreated [8].

Cognitive Behavioral Therapy

The evolution in psychological treatment for panic-type disorders has been rapid during the past 15 years. Recent cognitive-behavioral treatment protocols for anxiety and panic disorder consist of a multifaceted set of interventions, supplemented by homework. The homework reinforces positive attitudes and behaviors; paper questionnaires are most often used to collect this data. The data is taken back to therapy, where the clinician then gives the patient feedback on the information collected. Often patients forget to do this homework and return with incomplete or falsified data, setting back the therapeutic process.

Kalm Application Overview

The Kalm application is not a new direction for anxiety intervention, but rather a new tool based on CBT therapeutic techniques. The tool offers an abundance of user data points that were previously unavailable to clinicians, as well as real-time feedback for the user. Cognitive behavioral therapy's interventions are aimed at influencing mood, emotional reaction and behavioral patterns through the focused practice of new ways of thinking and acting. Based on these tenants, Kalm's activities present a holistic approach to increase the users' self-efficacy in the management of their anxiety. In addition to education about maladaptive and positive behaviors, the application guides the user through activities designed for use during times of heightened anxiety. Additional components include:

- A heart rate monitor
- A music-based activity to slow the user's heart rate when anxious
- Games designed to reinforce positive behaviors
- Push notifications for the user to self-regulate their interactions with the application
- Functionality for users to email their log data to themselves or a support person
- An anxiety tracker for users to chart their moods alongside contributing environmental and internal factors

The feedback screens for each of these activities shape user experience, particularly the visualization for the anxiety tracker. These screens provide the information users need to accurately evaluate their performance over time. Our assumption is that the presentation of the data may not be enough to promote self-awareness or persuade the user to modify their thoughts and behaviors in a positive way. More embellishment may be necessary to evoke change.

Mobile Health Apps in Practice

As mobile computing becomes part of our daily lives, its technologies affect the way we access and use information – medical and otherwise. This is particularly true for the millennial generation, a population at home with technology, games and applications [13]. The use of mobile device technology for mental health education and skill practice is perfectly suited to this demographic. A 2010 Carlson Marketing research study reported that 73% percent of teens and 93% of adults ages 18-29 own cell phones in the U.S. [4]. Of the teens with Internet access on their devices, 31% percent reported getting health, dieting or physical fitness information online [4]. If these trends are any indication, these numbers will only rise with time. With this rise comes the necessity for new mobile interface requirements.

The feedback mechanisms of three mobile applications inform the design of Kalm. All of the applications were designed to collect and present health related data, with the intent of fostering positive change in the lives of users. Moody Me, Bant and Mood Map all approach persuasive design in novel ways.

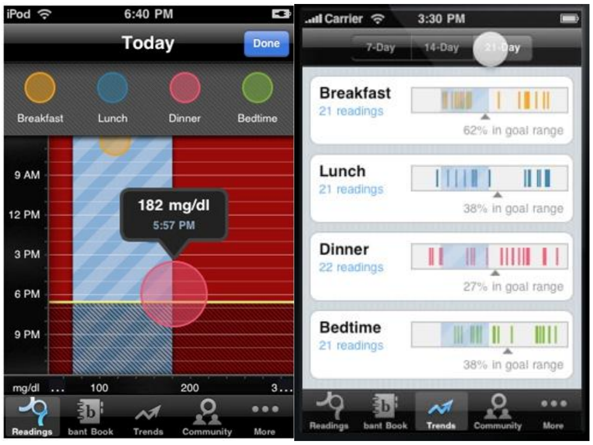
1.1.1 Bant

Created by the University Health Network and SickKids, Bant promotes health management for diabetic adolescents. The iPhone application supports teens in accurate decision making, helping them make medical adjustments for optimal health [11]. Teens can customize their inputs to categorize readings according to their activities and daily schedules. They earn experience points within the application that translate to the real world – they are rewarded with iTunes redemption codes to purchase music and other apps. Preliminary research on Bant has been very positive; research currently underway should shed light on mobile management of chronic disease with a youth population [11].

The two screens show here are for data input and feedback (figure 1). On the readings screen users select the meal that they want to do a glucose reading for. They then drag the colored dot associated to that meal into the graph area, which charts time of day against the glucose concentration of the meal. If the concentration hits the user-determined optimal range, the dot will fall within the blue section. Bant shows a graphical summary of these data points along their concentration ranges over time in the trends screen.

The Bant interface summarizes the data points well and allows for a good deal of organizational customization. Deviations from optimal are easy to find at a glance; it also clearly denotes fluctuation trends in the glucose readings. The information provided relies too heavy on color to differentiate between data points; it would be impossible for a color-blind, low-sighted, or blind user to use [7].

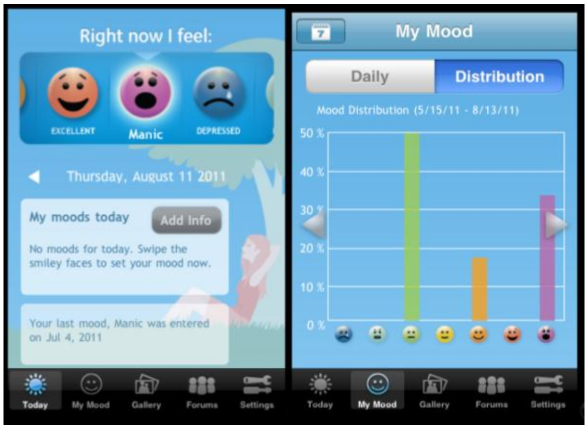
Figure 1. Bant Interface for readings (data collection) and trends (summary) [3]



1.1.2 Moody Me

Like Bant, Moody Me is also available through the iTunes store. The input screen is also the splash screen; the user swipes horizontally to find the smiley face that corresponds best to their current mood. The input screen tells the user what their average mood was for the past 7 days; for information on a greater period of time one must access the mood distribution graphs. The graphs chart the 7 mood options against a percentage scale and time.

Figure 2. Input and summary screenshots from Moody Me



No moods for today. Swipe the smiley faces to set your mood now.

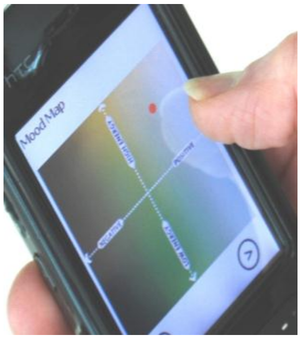
Your last mood, Manic was entered on Jul 4, 2011

The smiley faces in Moody Me are a major departure from most applications that rely solely on numerical graphs for feedback. The faces are easy to understand and speed data input. While the user may change the symptomology related to the faces, they may not change the actual faces or labels. Moody Me does not allow users to share their information with a clinician. Even if the functionality was included, it would be difficult for a clinician to get relevant data from a summary lacking numerical values (figure 2, left). Moody Me is most useful for the user desiring short term correlations or summaries.

1.1.3 Mood Map

While not commercially available, Mood Map was part of a qualitative exploration of how people adopt mobile therapies. Like Kalm, the Mood Map prototype offers functionality for users to track their moods in addition to exercises to practice relaxation techniques and positive behaviors. During the one month field study the application prompted users to log their moods several times a day, using randomized push notifications. While the study's focus wasn't design it did use user centered design techniques; it also shed light on user's willingness to use mobile therapies, which was promising.

Figure 1. Mood Map



1.2 The Interface of Kalm

We aim to create an interface that is intuitive and accessible to differently-abled users. Our definition of intuitive use – the subconscious application of prior knowledge – is defined by the German Intuitive Use of User Interfaces group [10]. An intuitive application results in: a low subjective mental workload, high perceived achievement of goals, low perceived effort of learning, high familiarity based on prior knowledge and a low perceived rate of error. Based on this definition the QUESI, questionnaire for the subjective consequences of intuitive use, will be used to evaluate the intuitive qualities of the application.

It is important to note that the evaluation of intuitiveness of the interface is not a measure of the actual content of the application. The QUESI uses a 5 level Likert-type scale but focuses on the consequence of use, rather than the features of the application. All of the questions are phrased in a way that higher scores represent a higher probability of intuitive use for example, "I could use the system without thinking about it" or "the system helped me to completely achieve my goals."

1.2.1 Anxiety/Experience Tracker

The reporting scales used within Kalm are based on the Multicenter Panic Anxiety Scale (MC-PAS, formerly the CY-PAS), the Sheehan Disability Scale, and the Beck Depression Inventory. The MC-PAS includes a four point scale that rates panic frequency and intensity, anticipatory anxiety, avoidance of sensations and situations, and impairment in work and social functioning. The SDS is a four-item self-report measure of impairment and the BDI is a well-validated measure of depressive symptomology. The scales used within Kalm represent a blend of self-rated disability and clinician-rated impairment measurements. The data from these reporting scales will be relevant to both the user and clinician if used in conjunction with therapy.

Sound will be used to describe and reinforce the imagery and information used in the application. Since chart embellishments can prove useful for intelligibility and recollection, we will repeat visual and aural metaphors across all of the various activities of the application [14].

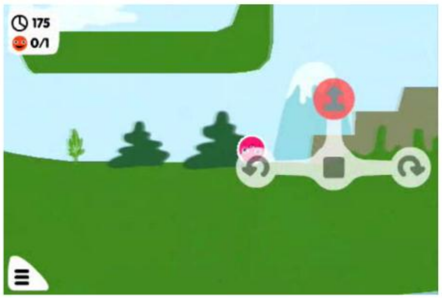
1.2.2 Games

Games are a good way to practice CBT homework while in a safe environment, particularly non-twitch based games (those requiring fast reflexes). The visual metaphors used in game can reinforce the information learned in the rest of the application. When gameplay incorporates a positive feedback loop, the experience becomes motivation to return to both the application and the game itself.

Game devices using a joystick for navigation often highlight associated menus and icons, subconsciously focusing the user's gaze to the most relevant icon position. Touch-based devices do not need a cursor in this capacity, since icons are acted on by touch. The navigation of Kalm's games present a special challenge due to a small amount screen space for gameplay and information.

The games for Kalm are informed by a touch scheme similar to that used in the game Rolando. Developed by Handcircus, Rolando is a platform game with puzzle elements that circumvents the iPhone's lack of screen space. All of the movements required by the game can be performed by using one's thumb; the interface then highlights all possible movement directions on the screen.

Figure 3. Rolando gameplay interface



The games will use little to no text; instead repeating the graphical icons and sound cues used throughout the other activities.

Discussion & Future Work

The smartphone is a viable venue for the presentation and practice of psychological health information. The creation of Kalm follows methodology conceived to support mobile interaction design [10, 5]. The use of Kalm returns a richer, larger data set than CBT information traditionally gathered with pen and paper. The feedback given by the application influences how users experience and perceive their own anxiety, so care must be taken when designing the interface and visual analytics of the system. Additionally, clinician input is necessary to identify counter-therapeutic practices or behaviors created by the use of the application.

Kalm is a tool designed to support patient accomplishment of self-management goals; beyond that it could be used as a platform for communication and collaboration between user and clinician. By providing an inexpensive, downloadable application, we believe that the number of individuals with access to CBT methodology will increase substantially. This is important not only for suffers but for any entity facing rising health care costs and barriers to providing quality mental health care.

REFERENCES

[1] Anderson, I., Maitland, J., Sherwood, S., Barkhuus, Chalmers, M., Hall, M., Brown, B. and Muller, H. Shakra: Tracking and Sharing Daily Activity Levels with Unaugmented Mobile Phones. Mobile Networks and Applications, 12 (2). 185-199.

[2] Arch gen psychiatry -- Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication, June 2005, Kessler et al. 62 (6): 593 Retrieved 2/1/2011, from http://archpsyc.ama-assn.org/cgi/content/full/62/6/593

[3] Bant Reading and Trends pages. Retrieved 8/30/11 from http://www.hospitalsongs.com/-wpcontent/uploads/2010/04/bant.jpg

[4] "Carlson Marketing Research Shows Segmenting Mobile Users Leads to Better Understanding and Greater Marketing Success." The Free Library 09 June 2010. Retrieved 1/4/2011, from <http://www.thefreelibrary.com/Carlson Marketing Research Shows Segmenting Mobile Users Leads to...-a0228452540>.

[5] Carriço, L., Zurita, G., Sá, M.D., Baloian, N., Faria, J., and Sá, I. Evaluating a Prototype for Geo-referenced Collaborative Psychotherapy with Mobile Devices. In Proceedings of CRIWG. 2010, 353-362.

[6] Dugas, M. J., Brillon, P., Savard, P., Turcotte, J., Gaudet, A., Ladouceur, R., et al. (2010). A Randomized Clinical Trial of Cognitive-Behavioral Therapy and Applied Relaxation for Adults with Generalized Anxiety Disorder. Behavior Therapy, 41(1), 46-58.

[7] Flatla, D., Gutwin, C. 2010. Individual Models of Color Differentiation to Improve Interpretability of Information Visualization. In *ACM Conference on Human Factors in Computing Systems (CHI 2010)*, Atlanta, Georgia, USA. 2563-2572.

[8] James, A., Soler, A., & Weatherall, R. (2005). Cognitive Behavioural Therapy for Anxiety Disorders in Children and Adolescents. Cochrane Database of Systematic Reviews (Online), (4), CD004690.

[9] Kessler RC, Berglund PA, Demler O, Jin R, Walters EE. Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV disorders in the National Comorbidity Survey Replication (NCS-R). Archives of General Psychiatry. 2005 Jun; 62(6):593-602.

[10] Mohs, C., Hurtienne, J., Israel, J. H., Naumann, A., Kindsmüller, M. C., Meyer, H. A., et al. IUUI–Intuitive Use of User Interfaces. In Bosenick, T., Hassenzahl, M., Müller-Prove, M., and Peissner, M. (Eds.), *Usability Professionals 2006*. German Chapter der Usability Professionals' Association, Stuttgart, 2006, 130-133.

[11] Murrary, Janice. 2011, Feb. 10. Bant iphone app to help teens manage type 1 diabetes [Web log message]. Retrieved 8/30/11 from http://www.newswire.ca/en/releases/archive/February2011/10/c9686.html

[12] Naumann, A., Hurtienne, J., Israel, J. H., Mohs, C., Kindsmüller, M. C., Meyer, H. A., and Hußlein, S. Intuitive Use of User Interfaces: Defining a vague concept. In Harris, D. (Ed.), Engineering Psychology and Cognitive *Ergonomics*. Springer, Heidelberg, 2007, 128-136.

[13] Overview | Pew Internet & American Life Project Retrieved 12/11/2010, from http://pewinternet.org/Reports/2010/Social-Media-and-Young-Adults/Summary-of-Findings.aspx

[14] Scott Bateman, Regan L. Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. 2010. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In Proceedings of the 28th international conference on Human factors in computing systems (CHI '10). ACM, New York, NY, USA, 2573-2582.

VAHC 2011

DEMOS

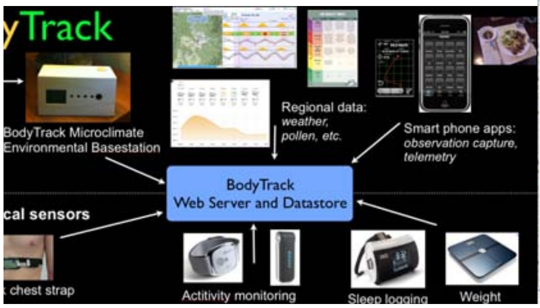
BodyTrack: Open Source Tools for Health Empowerment through Self-Tracking

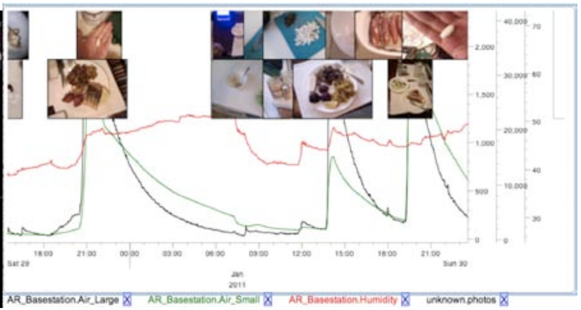
Anne Wright and Ray Yun, BodyTrack Project, CREATE Lab, Carnegie Mellon University

The BodyTrack project has interviewed a number of people who have improved their health by discovering certain foods or environmental exposures to avoid, or learning other types of behavioral changes. Many describe greatly improved quality of life, overcoming in some cases chronic problems in areas such as sleep, pain, gastrointestinal function, and energy levels. In some cases, a doctor or specialist's diagnosis led to treatment which mitigated symptoms (e.g. asthma or migraine headache), but where discovery of triggers required self-tracking and self-experimentation. Importantly, the act of starting to search for one's sensitivites or triggers appears to be empowering: people who embarked on this path changed their relationship to their health situation even before making the discoveries that helped lead to symptom improvement.

The BodyTrack Project is building tools, both technological and cultural, to empower more people to embrace an "investigator" role in their own lives. The core of the BodyTrack system is an open source web service which allows users to aggregate, visualize, and analyze data from a myriad of sources -- physiological metrics from wearable sensors, image and self-observation capture from smart phones, local environmental measures such as bedroom light and noise levels and in-house air quality monitoring, and regional environmental measures such as pollen/mold counts and air particulates. We believe empowering a broader set of people with these tools will help individuals and medical practitioners alike to better address health conditions with complex environmental or behavioral components.

We propose to demonstrate the current status of the BodyTrack tools and describe how we are using these tools to empower individuals to more powerfully explore their situation, and to enhance their ability to communicate and collaborate with their doctors and other health care providers.


yTrack
BodyTrack Microclimate Environmental Basestation
Regional data: weather, pollen, etc.
Smart phone apps: observation capture, telemetry
BodyTrack Web Server and Datastore
cal sensors
chest strap
Actitivity monitoring
Sleep logging
Weight



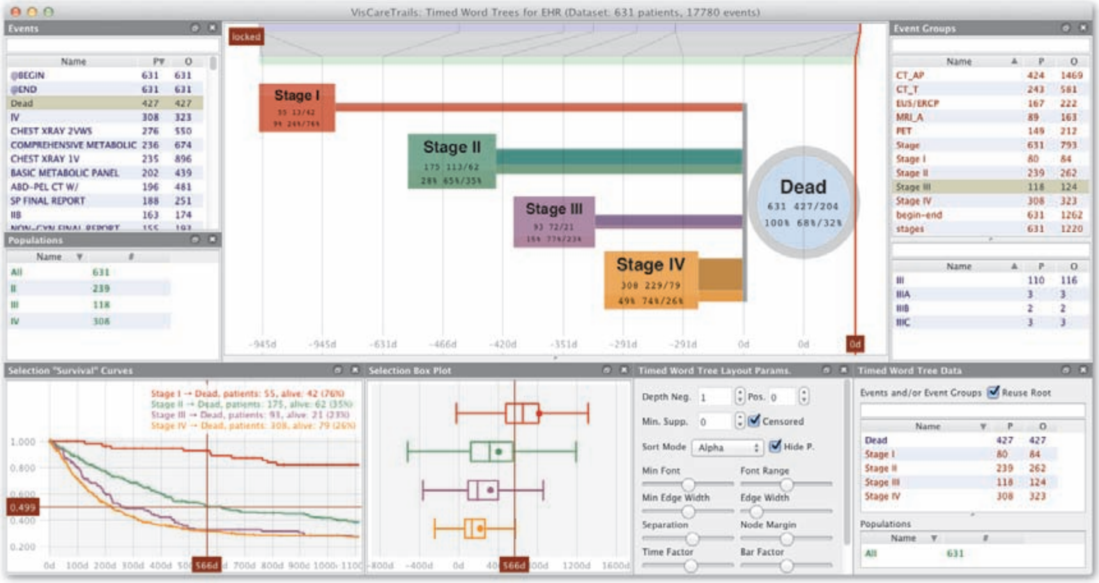
Demo of VisCareTrails: Visualizing Trails in the Electronic Health Record with Timed Word Trees, a Pancreas Cancer Use Case

Lauro Lins, Marta Heilbrun, Juliana Freire, and Claudio Silva

October 1, 2011

In this session we will demonstrate VisCareTrails, a system to visualize aspects of event sequences datasets (*e.g.,* set of patient histories). VisCareTrails features as its central display a visualization called *Timed Word Trees*, a generalization of Word Trees. The dataset we will use to demonstrate VisCareTrails consists of health care events on pancreatic cancer patients.

VisCareTrails supports the following pipeline: (1) a set of time-stamped event sequences is loaded into the system; (2) *group-events* are defined as needed (Stage III in Figure 1 is a group-event that means either event III, IIIA, IIIB or IIIC); (3) a timed word tree is generated by dragging and dropping events and/or group-events into the central canvas (in Figure 1, stage events & Dead were dragged and dropped into the canvas); (4) one of the dropped events is defined as the root event (by default the root is the first element that was dropped in the visualization, but a user can change the root event at any time); (5) the visual summary generated is inspected to understand paths that end and start in the root event; and (6) path nodes are selected to obtain survival curves for the sequences. Figure 1 shows survival curves of the selected stage nodes (red, green, purple, and orange paths): bottom left widget. The visual summary conveys information about frequency of events (larger fonts and thicker transitions means more sequences going through the path), time distances (based on average times) of the events relative to their parent event; and a hint on the dispersion (*i.e.,* standard deviation) of time distances in each event transition (i.e. the hue of blue darkens as the standard deviation of the time distance decreases). On the second bottom widget (from left to right), we show a box-plot for the time distance distribution from the selected events to the root event.



VisCareTrails: Timed Word Trees for EHR (Dataset: 631 patients, 17780 events)

Events

Name	P▼	O
@BEGIN	631	631
@END	631	631
Dead	427	427
IV	308	323
CHEST XRAY 2VWS	276	550
COMPREHENSIVE METABOLIC	236	674
CHEST XRAY 1V	235	896
BASIC METABOLIC PANEL	202	439
ABD-PEL CT W/	196	481
SP FINAL REPORT	188	251
IIB	163	174
NON-GYN FINAL REPORT	155	193

Populations

Name ▼	#
All	631
II	239
III	118
IV	308

Event Groups

Name ▲	P	O
CT_AP	424	1469
CT_T	243	581
EUS/ERCP	167	222
MRI_A	89	163
PET	149	212
Stage	631	793
Stage I	80	84
Stage II	239	262
Stage III	118	124
Stage IV	308	323
begin-end	631	1262
stages	631	1220

Name ▲	P	O
III	110	116
IIIA	3	3
IIIB	2	2
IIIC	3	3

Selection "Survival" Curves

Stage I → Dead, patients: 55, alive: 42 (76%)
Stage II → Dead, patients: 175, alive: 62 (35%)
Stage III → Dead, patients: 93, alive: 21 (23%)
Stage IV → Dead, patients: 308, alive: 79 (26%)

Selection Box Plot

Timed Word Tree Layout Params.

Depth Neg. 1 Pos. 0
Min. Supp. 0 ☑ Censored
Sort Mode Alpha ☑ Hide P.

Timed Word Tree Data

Events and/or Event Groups ☑ Reuse Root

Name ▼	P	O
Dead	427	427
Stage I	80	84
Stage II	239	262
Stage III	118	124
Stage IV	308	323

Populations

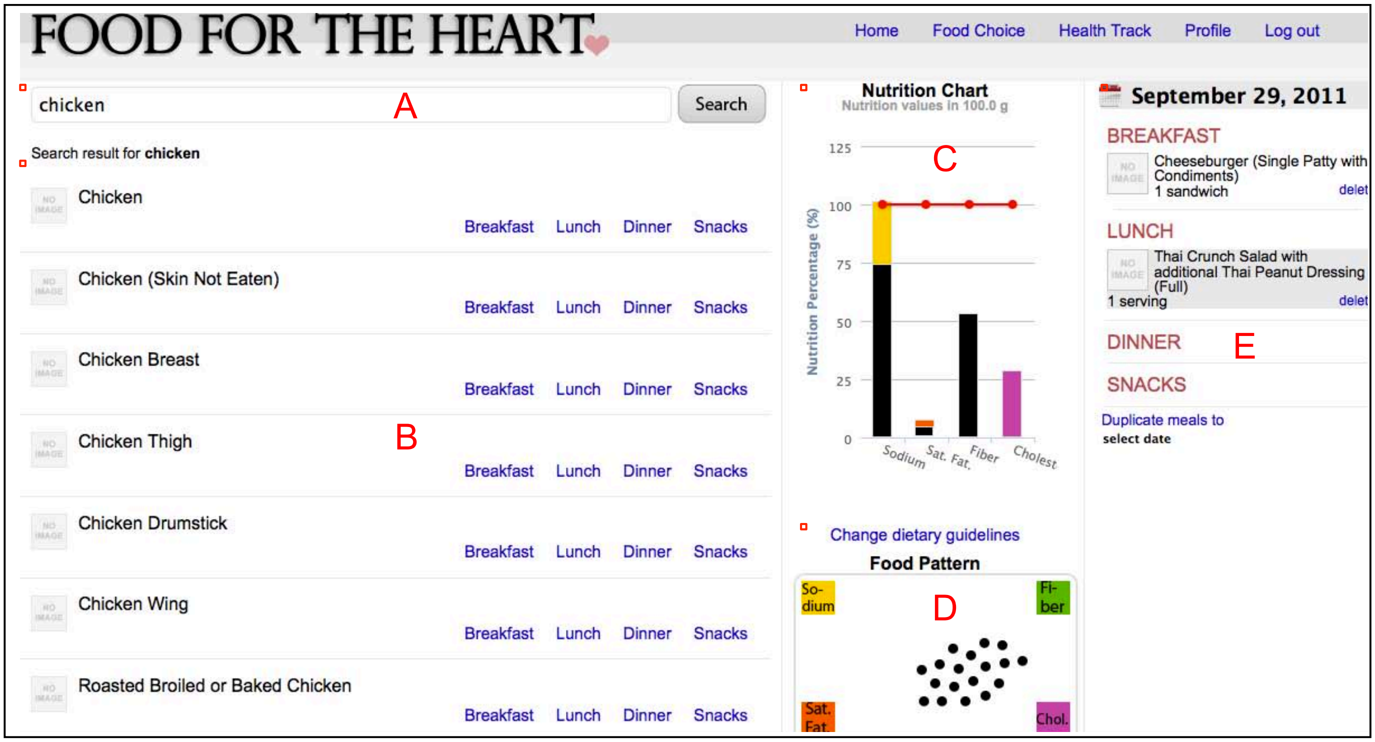
Name ▼	#
All	631

Figure 1: VisCareTrails session on a dataset of pancreatic cancer patients

Food For The Heart: Visualizing Nutritional Contents for Food Items for Patients with Coronary Heart Disease

Fransisca Vina Zerlina[1], Bum chul Kwon[2], Sung-Hee Kim[2],
Karen S. Yehle[2], Kimberly S. Plake[2], Sibylle Kranz[2], Lane M. Yahiro[2], and Ji Soo Yi[2]

University of Washington[1] Purdue University[2]



1 Introduction

We introduce a web-based interactive nutrition-based food selection tool for patients with Coronary Heart Disease (CHD), called "Food ForThe Heart (FFH)," inspired by a multivariate information visualization tool, called Dust & Magnet [1]. FFH visualizes four core nutritional components (sodium, fiber, saturated fat, and cholesterol) of more than 30,000 food items in two different visualizations, a bar chart and the Dust & Magnet view, upon user's request. In particular, The Dust & Magnet view provides an overview of multiple food items based on the four core nutritional components in a two-dimensional pane so that a user can easily find which food items are suitable for their diet.

2 Design Details

FFH is a web-based dietary intervention system whereby a user can evaluate multiple food items based on their nutritional content. The main page is designed as shown at the top. It consists of five segments labeled them from A to E with red borderlines for convenience. Each segment is described below :

- A (Search Box): A user can enter the name of a food item (e.g., chicken) or search by brand names (e.g., McDonald's).
- B (Search Results): A list of food items generated by an entered search query. The nutritional content of each item can be found in two graphs, C and D, upon the user's click. Each item could be inserted into one of the meal boxes in E.
- C (Nutrition Chart): The chart visualizes accumulated nutrition values of food items selected from B across the four nutritional components (sodium, saturated fat, fiber, and cholesterol) per day. The Y-axis indicates the percentage of the recommended amounts (100%) of each nutritional component.
- D (Dust & Magnet): Using the dust particles (food items) and magnets (four nutritional components) metaphor, the system spreads out food items based on the amount of each nutritional component. For example, the more sodium in a food item, the closer it is positioned to the sodium magnet.
- E (Diet Plan): Food items inserted from B can be seen here. A user can also view dietary intake from previous days by switching dates.

3 Demonstration

The demonstration will walk through how a user can control his/her diet plan by using this website. The audience will participate in using the system on the fly by adding their favorite food items to see how healthy their choices are. The demonstration will provide insights about how we can help patients control their diet by leveraging the power of visualization systems.

References

[1] Yi, J. S., Melton, R., Stasko, J., & Jacko, J. A. (2005). Dust & Magnet: Multivariate Information Visualization using a Magnet Metaphor. Information Visualization, 4(4), 239-256.

ImageVis3D Mobile in Clinical Use

Jens Krüger[*] Thomas Fogal[†]

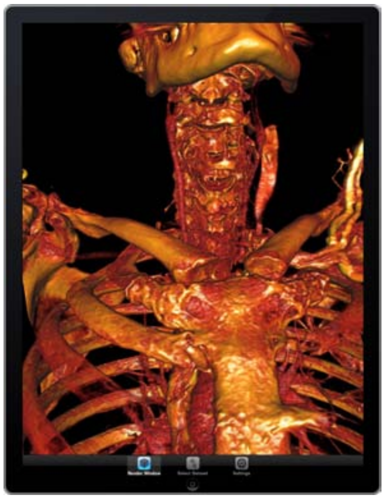


Figure 1: ImageVis3D Mobile displaying a contrast-enhanced CT scan of a human male. High-quality imagery can be streamed from more capable rendering resources, or rendered directly on the device.

1 Introduction

While there have been a variety of research applications developed for mobile platforms [2, 1, 6, 5, 8, 9], there has been less work in applying such techniques in a clinical setting. A notable exception is Meir and Rubinsky's work [7], which attempted to use a mobile system to improve cancer diagnosis. Despite this, there has been no publicly available open platform for the deployment of mobile medical visualization systems.

ImageVis3D Mobile ("IV3Dm") is a visualization application for mobile devices, enabling physicians to bring clinical practice to the point of care. Using IV3Dm, visual feedback can be disseminated to trained professionals to aid them in interpreting data, feedback from other physicians can be obtained in chance 'hallway meetings', and patient data can be communicated directly to the patient in a manner they can interpret and understand.

2 From Supercomputer to Tablet

ImageVis3D Mobile and its desktop counterpart, ImageVis3D, run on supercomputers and tablets, as well as everything in between [3]. Using the desktop version of ImageVis3D, one can visualize data of unlimited size using commodity workstations [4]. This data-size-agnostic feature has become critical as modern scanners continue to produce scans which are a challenge to view in anything but two dimensions. ImageVis3D allows high resolution three-dimensional reconstructions of CT and MRI data (as well as a variety of other data types), in an easy to use, lightweight application. We continue to expand our exploration of extensive computing installations to render larger data in real time.

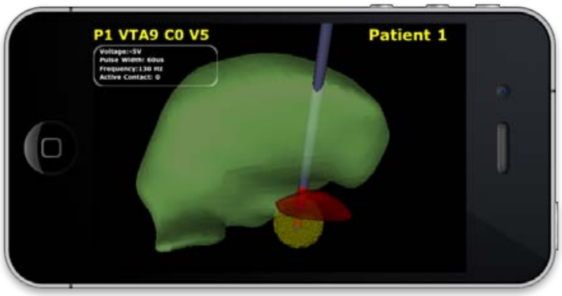


Figure 2: Initial application of ImageVis3D Mobile: providing visual feedback for setting deep brain stimulation parameters.

However, we believe the revolutionary aspect of this software system is in its ability to visualize one's data on mobile devices. This software enables physicians to look at data in an entirely new dimension, uncovering aspects which have never been visible before. An 'always on' device for medical visualization opens up new avenues for collaboration and data dissemination which are not possible in a more traditional clinical setting.

3 Zero Infrastructure

A common problem with novel medical visualization techniques is transferring them from a research environment to clinical practice. Scalable infrastructures must be created and maintained, entailing laborious procedural red tape and excessive per-client configuration. With ImageVis3D and ImageVis3D Mobile, these boundaries are quickly being broken down. We are developing a cloud-based infrastructure to enable new research groups to utilize the system with 0 configuration. Since the client software is open source and freely available, installation is as simple as downloading any other application. Utilizing platforms that physicians own personally and already carry with them regularly ensures that there is no administrative overhead to applying the system.

4 Effective in Practice

Our first application of IV3Dm in a clinical setting has been in the area of deep brain stimulation (DBS), more specifically for the treatment of Parkinson's disease. For the process to be successful, a DBS planning 'programmer' must work with the patient to provide individualized stimulation parameters. This can be a lengthy process using the abstract 'set and test' method which is the standard of care, but using the visual IV3Dm platform as shown in Figure 2, clinical support staff can perform the operation an order of magnitude more quickly.

[*] IVDA, DFKI, Intel VCI, SCI

[†] SCI

References

[1] S. Burigat and L. Chittaro. Location-aware visualization of vrml models in gps-based mobile guides. 2005.

[2] C. Chang and S. Ger. Enhancing 3D graphics on mobile devices by image-based rendering. 2002.

[3] Thomas Fogal, Hank Childs, Siddharth Shankar, J. Krüger, R. D. Bergeron, and P. Hatcher. Large Data Visualization on Distributed Memory Multi-GPU Clusters. 2010.

[4] Thomas Fogal and Jens Krüger. Tuvok - an architecture for large scale volume rendering. 2010.

[5] F. Lamberti and A. Sanna. A Streaming-Based Solution for Remote Visualization of 3D Graphics on Mobile Devices. 2007.

[6] J. Lluch, R. Gaitán, E. Camahort, and R. Vivó. Interactive three-dimensional rendering on mobile computer devices. 2005.

[7] A. Meir and B. Rubinsky. Distributed network, wireless and cloud computing enabled 3-D ultrasound; a new medical technology paradigm. 2009.

[8] M. Moser and D. Weiskopf. Interactive Volume Rendering on Mobile Devices. 2008.

[9] S. Park, W. Kim, and I. Ihm. Mobile collaborative medical display system. 2008.

DEMO:
AnamneVis: A Framework for the Visualization of Patient History and Medical Diagnostics Chains

Zhiyuan Zhang[1], Faisal Ahmed[1], Arunesh Mittal[1], IV Ramakrishnan[1], Rong Zhao[1], Asa Viccellio[2], and Klaus Mueller[1]

[1]Computer Science Department and Center for Wireless and Information Technology (CEWIT)

[2]Department of Emergency Medicine

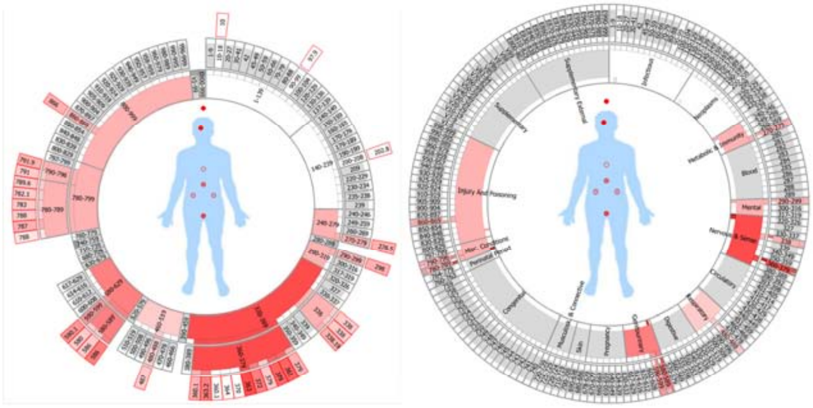
Stony Brook University

The medical history or *anamnesis* of a patient is the factual information obtained by a physician for the medical diagnostics of a patient. This information includes current symptoms, history of present illness, previous treatments, available data, current medications, past history, family history, and others. Based on this information the physician follows through a medical diagnostics chain that includes requests for further data, diagnosis, treatment, follow-up, and eventually a report of treatment outcome. Patients often have rather complex medical histories, and visualization and visual analytics can offer large benefits for the navigation and reasoning with this information. Here we will demo *AnamneVis*, a system where the patient is represented as a radial sunburst visualization that captures all health conditions of the past and present to serve as a quick overview to the interrogating physician. The patient's body is represented as a stylized body map that can be zoomed into for further anatomical detail. On the other hand, the reasoning chain is represented as a multi-stage flow chart, composed of date, symptom, data, diagnosis, treatment, and outcome.

Our health care informatics prototype aims to provide a comprehensive multi-faceted assessment of the patient and his (her) history for intuitive information retrieval by the physician. The goal is information organization and integration along these various aspects. Overview and detail-on-demand requires hierarchies, and effective information organization requires robust encoding by ways of well-established criteria – we use standard codes commonly used for billing in hospitals which enables us to easily build our system on top of an existing health care information system. These codes are ICD, CPT, and NDC. ICD is the code used to describe the condition or disease being treated, also known as the diagnosis. CPT is the code used to describe medical services and procedures performed by doctors for a particular diagnosis. NDC is the code used for administered drugs. Further goals, often expressed by our collaborating emergency physician are ease of information access and flexibility in displayed aggregated information and data. To enable this functionality, our system is fully interactive and the displays are fully linked and coordinated. In the following we show snapshots of these displays.

The hierarchical radial display is used primarily to show information about the patient. There are three cooperating displays: (1) symptoms and diagnoses, (2) procedures and treatments, (3) data. These three displays are interlinked to allow doctors to obtain a full picture of the patient as well as assess existing relationships. Two examples of this display are shown in Fig. 1.

The sequential display is used mainly to demonstrate the medical diagnostic flow. The medical records are organized by an underlying graph data structure. Each node corresponds to one incident (medical primitive), which could be a doctor visit, symptom, test/data, diagnosis or treatment. Edges represent relationships. An example of this display is shown in Fig. 2.



(a) Hierarchy-centric layout (b) With ICD code name/label overlay

Figure 1. Sunburst display for patient information.

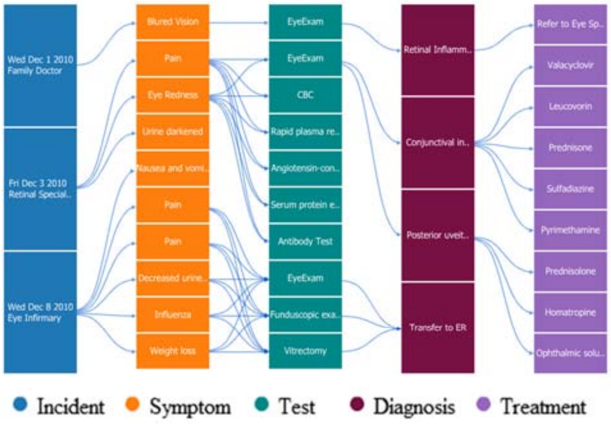
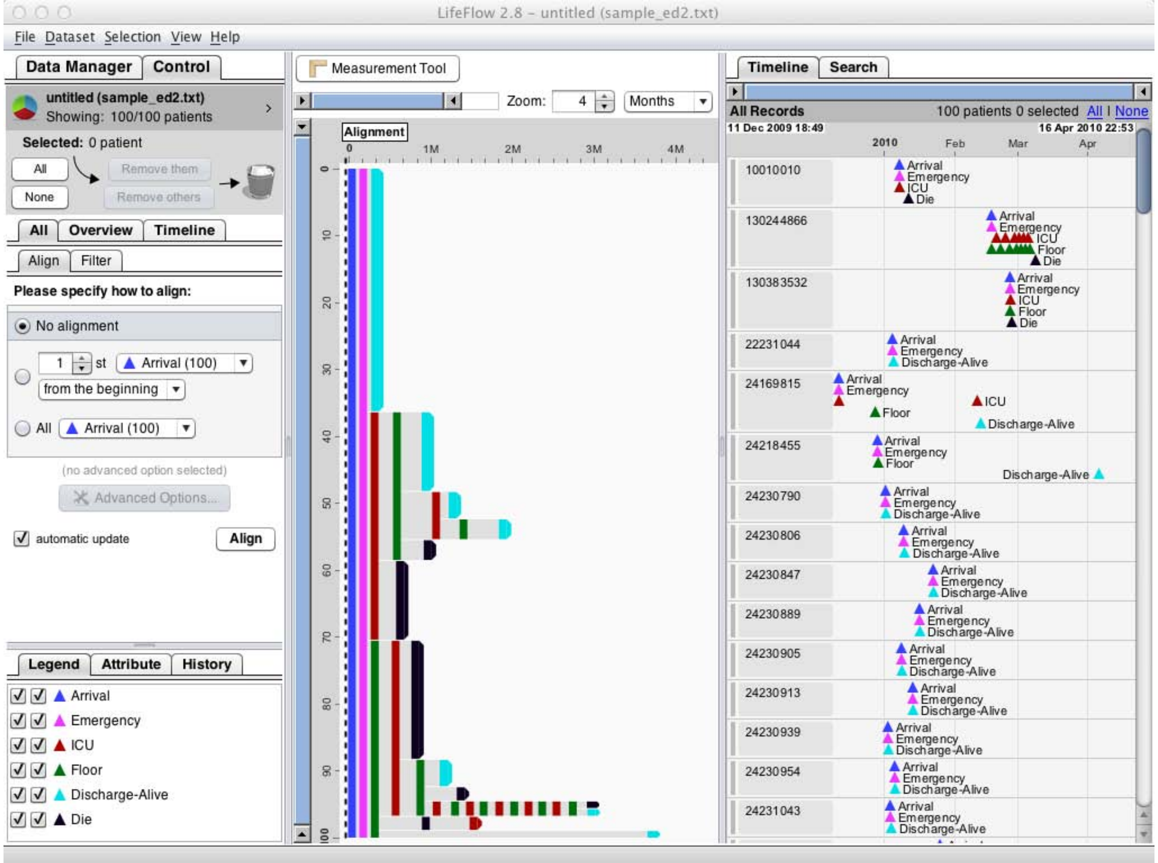


Figure 2. Sequential display for medical diagnostics chain.

Acknowledgments: We would like to express our deepest thanks to the UC Berkeley Visualization Lab and Stanford HCI group to make available the very helpful Flare toolkit, which we used as a basis and extended non-trivially for the system presented in this work.

LifeFlow: Understanding Millions of Event Sequences in a Million Pixels

Krist Wongsuphasawat



LifeFlow 2.8 – untitled (sample_ed2.txt)

untitled (sample_ed2.txt)
Showing: 100/100 patients
Selected: 0 patient

Please specify how to align:
● No alignment
○ 1 st Arrival (100) from the beginning
○ All Arrival (100)
(no advanced option selected)
☑ automatic update

Zoom: 4 Months

All Records 100 patients 0 selected All I None
11 Dec 2009 18:49 – 16 Apr 2010 22:53

File
Dataset
Selection
View
Help
Data Manager
Control
All
None
Remove them
Remove others
All
Overview
Timeline
Align
Filter
Advanced Options...
Align
Legend
Attribute
History
☑ ☑ Arrival
☑ ☑ Emergency
☑ ☑ ICU
☑ ☑ Floor
☑ ☑ Discharge-Alive
☑ ☑ Die
Measurement Tool
Alignment
0
1M
2M
3M
4M
0
10
20
30
40
50
60
70
80
90
100
Timeline
Search
2010
Feb
Mar
Apr
10010010 Arrival Emergency ICU Die
130244866 Arrival Emergency ICU Floor Die
130383532 Arrival Emergency ICU Floor Die
22231044 Arrival Emergency Discharge-Alive
24169815 Arrival Emergency Floor ICU Discharge-Alive
24218455 Arrival Emergency Floor Discharge-Alive
24230790 Arrival Emergency Discharge-Alive
24230806 Arrival Emergency Discharge-Alive
24230847 Arrival Emergency Discharge-Alive
24230889 Arrival Emergency Discharge-Alive
24230905 Arrival Emergency Discharge-Alive
24230913 Arrival Emergency Discharge-Alive
24230939 Arrival Emergency Discharge-Alive
24230954 Arrival Emergency Discharge-Alive
24231043 Arrival Emergency Discharge-Alive

Description

Event sequence analysis is an important task in many domains: medical researchers study the patterns of transfers within the hospital for quality control; transportation experts study accident response logs to identify best practices. In most cases they deal with more than thousands of records. While previous research has focused on searching and browsing, overview tasks are often overlooked. We introduce a novel interactive visual overview of event sequences called *LifeFlow*. LifeFlow scales to any number of records, summarizes all possible sequences, and highlights the temporal spacing of the events within sequences.

Please visit http://www.cs.umd.edu/hcil/lifeflow for more details

Reference

Krist Wongsuphasawat, John Alexis Guerra Gómez, Catherine Plaisant, Taowei David Wang, Meirav Taieb-Maimon and Ben Shneiderman *LifeFlow: Visualizing an Overview of Event Sequences* In Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems (CHI'11), 1747-1756.

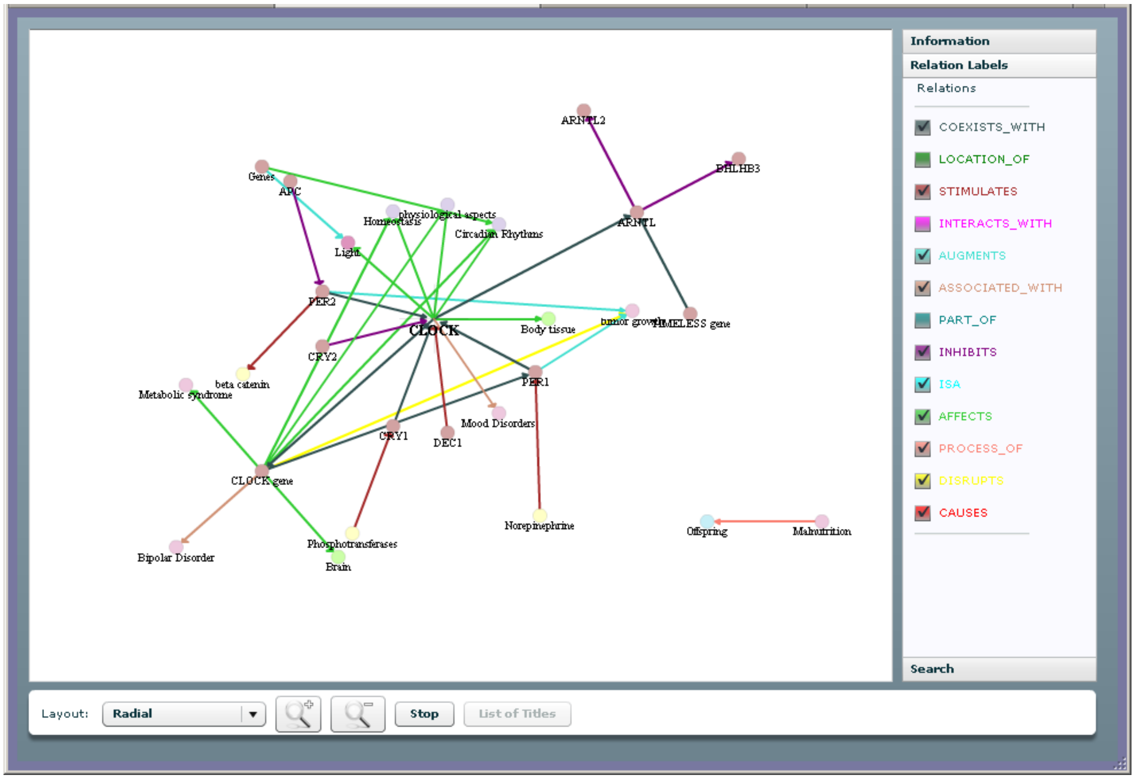
Title: Semantic MedLine

Authors: Michael J. Cairelli and Thomas C. Rindflesch

Affiliation: Lister Hill National Center for Biomedical Communications, National Library of Medicine

Contact: mike.cairelli@nih.gov, trindflesch@mail.nih.gov

We propose a demo of Semantic MEDLINE using an investigation of recent literature on clock genes to demonstrate the tool's ability to summarize recent literature. Clock genes were discovered in fruit flies in 1971 and were later found in all other organisms, including humans. Recent research has gone beyond molecular mechanics into the physiological implications of inappropriate expression and/or function of clock genes in common diseases with significant clinical implications. We will do a search for "clock genes," using the most recent 1000 citations and then summarize the results using the Pharmacogenomics schema to produce a graph of related predications. Then we will navigate through the graph, highlighting significant predications and the underlying articles that provided them. This tutorial will introduce the audience to Semantic MEDLINE's general capabilities, the advantages of its visual representation of search results, and some opportunities for tool development.



Layout: Radial

InBox: In-situ Multiple-Selection and Multiple-View Exploration of Diffusion Tensor MRI Visualization

Jian Chen [*] Haipeng Cai
University of Southern Mississippi

Alexander P. Auchus
University of Mississippi Medical Center

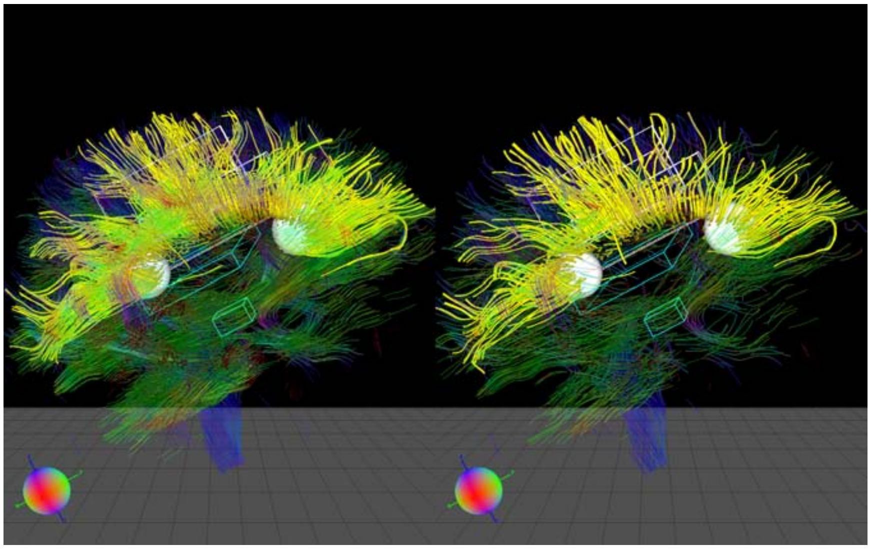


Figure 1: The InBox multiple coordinated view interface: the user can use *in-situ* box(es) and sphere(s) to select regular or angular shaped brain regions. Here selected fibers are within the white-colored boxes and spheres. Two removal boxes are colored in blue. Here, the corpus callosum (a wide, flat bundle of neural fibers connecting the left and right hemispheres) is selected by operating the widgets from one view to select fibers in dual views. The datasets are from a normal brain sampled at different seeding resolutions. The two streamtube models of human brains are from a normal brain model in voxel resolutions of $0.94mm \times 0.94mm \times 4.52mm$ (left) and $1.72mm \times 1.72mm \times 3.00mm$ (right) accordingly.

Abstract

We will demonstrate InBox, an *in-situ* multiple-selection and multiple-view interface for interactive exploration of dense tube-based diffusion tensor magnetic resonance imaging (DTI) visualization. DTI is an *in vivo* non-invasive technique that measures the directional dependence of the motion of water molecules in tissues in three-dimensions (3D). Fiber tracking or tractography is a standard approach to visualize the results of DTI. The tractography produces a set of integral curves or fibers that follow the principle direction of diffusion. If fibers are constructed and visualized individually through a large volume of DTI, the display gets cluttered making it difficult to get insight in the data. Thus, efficient interaction is often demanded.

A high-level contribution of InBox is the design considerations for the tight integration of selection with widget-based interface. Built on existing techniques and suggestions provided by our DTI collaborators, our work focuses on the use of conventional desktop setting and helps users *stay in the flow* of focused attention. This work builds on the assumption that focusing on the current working window can facilitate more precise selection by engaging the users in their tasks. We call our interface InBox to stand for *in-situ* selection.

Figure 1 shows a scenario of use of InBox, where the primary interface is box- and sphere-based for selecting regular and angular shapes. The box-based design was inspired by existing selection approaches in BrainApp [2] and CINCH [1]. The sphere-widget was added to select regions that include curved fibers to exclude unwanted ones that would have been chosen with a box. Box and sphere widgets are as freely added and deleted as needed.

A unique property was to provide interactive side-by-side views of the data across different scales that are designed to be less demanding on visual attention by enabling in place actions. It leverages the use of screen space for splitting actions. Multiple boxes or spheres can function together using two working modes (selection and removal) and two associative logics (AND and OR) to support a rich set of operations. Views are coordinated to facilitate tasks such as comparing a patient's DTI captured at different time instances. For example, our collaborators were interested in comparing the brain development of two cases (one normal and one agenesis of corpus callosum or ACC). What the doctor who used InBox did was to put the two datasets side by side, cull out those peripheral fiber bundles using the selection boxes, and then fully engage in the fiber bundles around CC. They did confirm their hypothesis using our visualization with the InBox interface. Such an interface can also be useful for educational purposes for showing cases to the medical school students.

References

[1] D. Akers. CINCH: a cooperatively designed marking interface for 3d pathway selection. ACM UIST, pages 33–42, New York, NY, USA, 2006. ACM.

[2] S. Zhang and D. H. Laidlaw. Visualizing diffusion tensor MR images using streamtubes and streamsurfaces. *IEEE TVCG*, 9(4):454–462, 2003.

[*] Contact: jian.chen@usm.edu

Program Committee

Organizers:

Jesus J Caban, PhD
NICoE / Naval Medical Center
CC / National Institutes of Health

David Gotz, PhD
IBM T.J. Watson Research Center

Program Committee	
Michael J. Ackerman, PhD	National Library of Medicine
Alark Joshi, PhD	Boise State University
Robert Kosara, PhD	University of North Carolina at Charlotte
Joseph Kvedar, MD, PhD	Center for Connected Health / Harvard
Paul G. Nagy, PhD	Department of Radiology, Johns Hopkins University
Adrian Park, MD	Department of Surgery, Dalhousie University
Catherine Plaisant, PhD	University of Maryland, College Park
Penny Rheingans, PhD	University of Maryland, UMBC
Terry S. Yoo	National Library of Medicine

Author Index

Jake Y. Chen	Indiana University School of Informatics	56
James J Cimino	National Institutes of Health	29
Jason Mclaughlin	Indiana University - Purdue University	56
Jens Kruger	University of Utah	69
Ji Soo Yi	Purdue University	68
Jian Chen	University of Southern Mississippi	74
Jim DeLeo	National Institutes of Health	29
Juliana Freire	NYU-Poly	13, 67
Karen S. Yehle	Purdue University	68
Kevin Hughes	University of Massachusetts Lowell	41
Kimberly S Plake	Purdue University	68
Klaus Mueller	Stony Brook University	17, 71
Kostas Pantazos	University of Copenhagen	21
Krist Wongsuphasawat	University of Maryland	25,72
Lane M Yahiro	Purdue University	68
Lauro Lins	NYU-Poly	13, 67
Marta Heilbrun	University of Utah	13, 67
Michael J Cairelli	National Library of Medicine	37, 73
Patrick Gillich	US Army Research Laboratory	53
Qian You	Indiana University - Purdue University	56
Raghu Machiraju	Ohio State University	33
Ray Yun	Carnegie Mellon University	66
Richard Arias-Hernandez	Simon Fraser University	45
Rong Zhao	Stony Brook University	17, 71