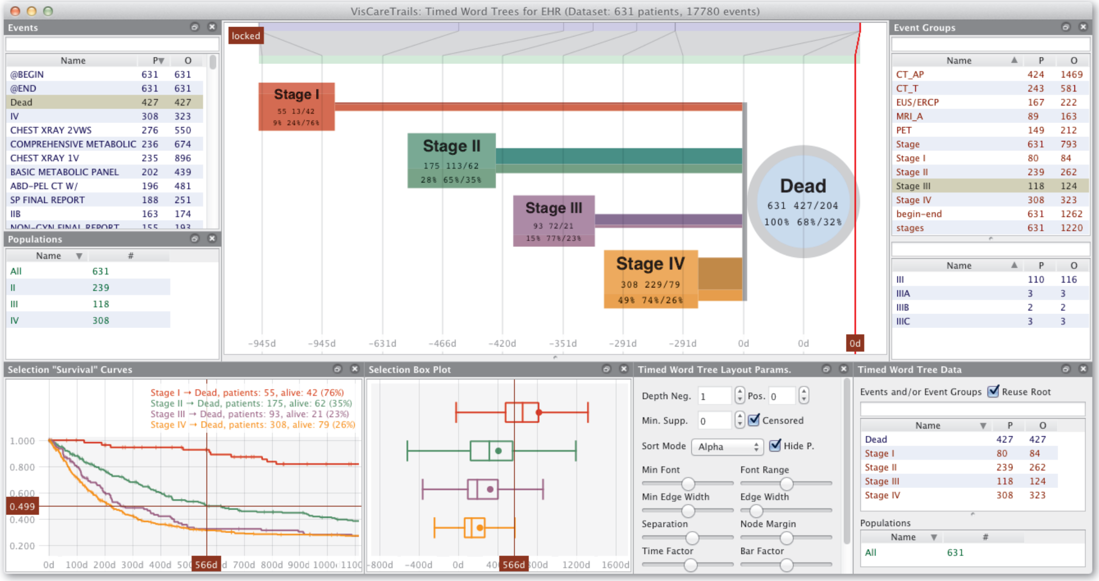
VISCARETRAILS: Visualizing Trails in the Electronic Health Record with Timed Word Trees, a Pancreas Cancer Use Case

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



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-CYN 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%)

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.

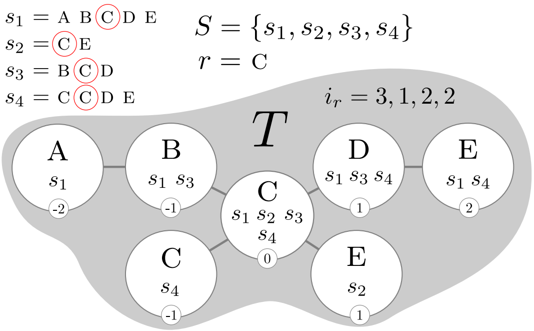


Fig. 2. Example of an event tree, T, rooted at event C for the set of event sequences S. Each node in T has an event label, a subset of sequences, and an offset (small circle). The central visualization in VisCareTrails are visual representations for event trees.

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

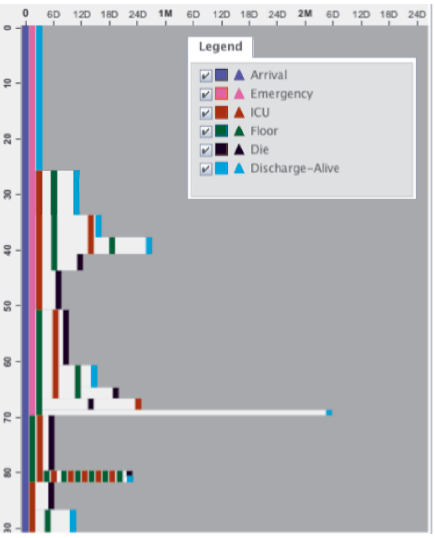


Fig. 3. LifeFlow visualization, V_{LF}, summarizing hospital event sequences for 91 patients (taken from [9]).

(*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.

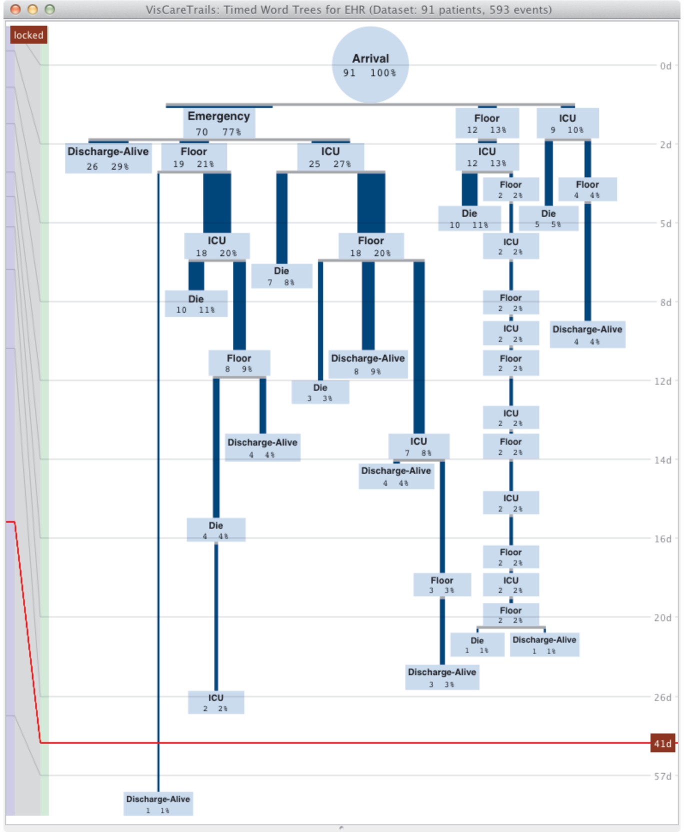
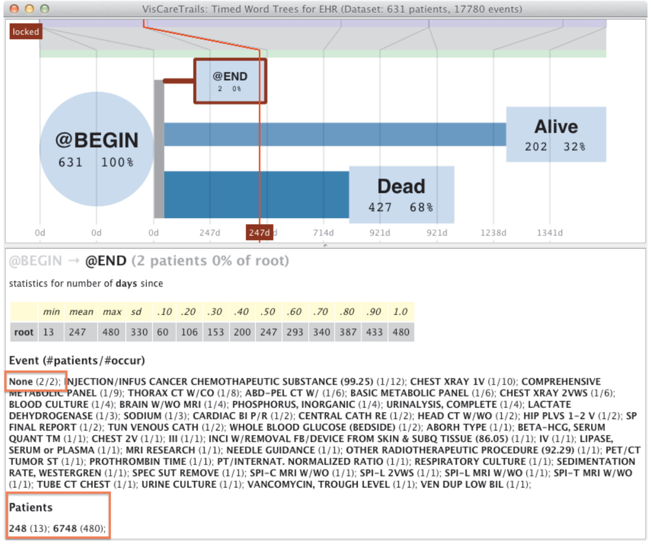


Fig. 4. Proposed timed word tree visualization, V_{TWT}, in VisCareTrails for the same event tree of Figure 3.

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.

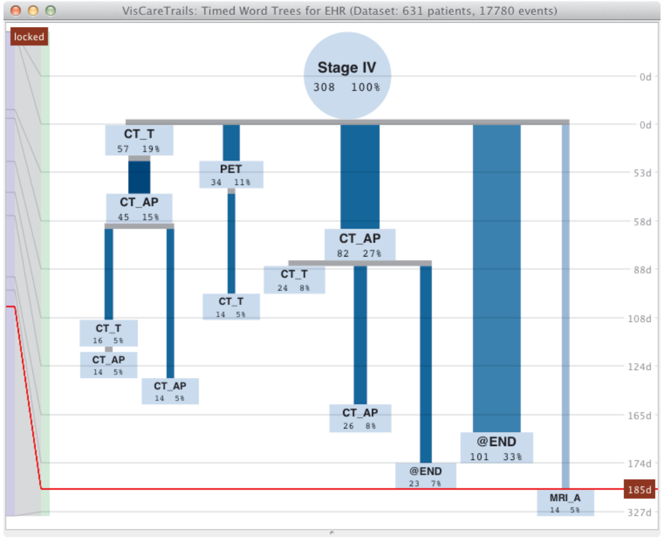


Fig. 6. Timed word tree with most frequent event paths (≥ 14 patients) after a patient gets registered in Stage IV. Events considered are in diagnostic test groups CT_AP, CT_T, EUS/ERCP, MRI_A, or PET. Event @END was included to indicate frequent paths where no event in a diagnostic test group occurred (*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.