Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records

Taowei David Wang, Catherine Plaisant, Alexander J. Quinn, Roman Stanchak, Ben Shneiderman
Department of Computer Science
University of Maryland, College Park, MD 20742
{tw7, plaisant, aq, roman, ben}@cs.umd.edu

Shawn Murphy
Massachusetts General Hospital
50 Staniford Street, 7th floor
Boston, MA 02114
murphy.shawn@mgh.harvard.edu

ABSTRACT
Electronic Health Records (EHRs) and other temporal databases contain hidden patterns that reveal important cause-and-effect phenomena. Finding these patterns is a challenge when using traditional query languages and tabular displays. We present an interactive visual tool that complements query formulation by providing operations to align, rank and filter the results, and to visualize estimates of the intervals of validity of the data. Display of patient histories aligned on sentinel events (such as a first heart attack) enables users to spot precursor, co-occurring, and aftereffect events. A controlled study demonstrates the benefits of providing alignment (with a 61% speed improvement for complex tasks). A qualitative study and interviews with medical professionals demonstrates that the interface can be learned quickly and seems to address their needs.

Author Keywords
Information visualization, evaluation, electronic health record, search, uncertainty, temporal data.

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces.

INTRODUCTION
Discovering patterns is a common step of scientific inquiry. Medical researchers are interested in temporal patterns across health records. For example, selecting patients for clinical trials requires careful review of patients with similar medical history. Previous work has shown that a timeline visualization for personal histories can provide benefits over a tabular view [16], but a timeline alone does not address all of the tasks users face. In particular, tasks that involve temporal comparisons relative to important events such as a heart attack are not supported. We explored strategies in supporting such tasks in our Lifelines2 project. In this paper we first motivate the problems from the medical domain, describe the interface, and discuss the results of our evaluations. While the examples in this paper are medical in nature, the type of temporal data analysis we address is common in other fields such as surveillance and intelligence [10], criminal activities patterns [5], Web session log analysis [11], or the study of human activities in general.

To appear in the proceedings of:
CHI 2008, April 5–10, 2008, Florence, Italy.

MOTIVATION
Our colleagues at Harvard Medical School highlight two scenarios where querying large databases of clinical data and reviewing results are not well addressed by currently available tools: 1) observational research using existing data (instead of a clinical trial). In this scenario, researchers use de-identified data collected for the purpose of other (existing) studies or medical practice data to better understand health problems or study the effect of treatments. 2) Clinical trial patient recruitment. Harvard receives over 600 requests a year from researchers everywhere to find suitable participants for clinical trials. The query terms typically contain diagnoses, treatments, and chief complaints. Analysts need to make a decision on whether each of the patients is a possible candidate for the trial based on manual reviews of the results. Among the many challenges posed by these two scenarios, two issues involve display design and user interaction. First, temporal comparison among patients is challenging, subsequently making finding patterns difficult. Secondly, even experts who know the underlying temporal data semantic and provenance can misinterpret the data presented. We illustrate these two challenges in the following example:

Researchers who study asthma may be interested in the relationship between patients' first pneumonia and their asthma attacks and treatments. A query is issued with those two diagnoses. Researchers then review the frequency and the temporal placement of the asthma events in relation to the first pneumonia in a large set of patient records. We call the reference event (i.e. the 1st occurrence of pneumonia) the *sentinel event*. Reviewing

results on a timeline display is helpful but requires comparing events spread over, potentially, a large time span. In addition, pneumonia and asthma are often related, and diagnoses often occur very closely, researchers may spend significant amount of time zooming and panning to switch between a global overview and a detailed inspection. The constant interaction with the display is disruptive to a researcher's visual memory, making discovery of patterns difficult. Proving the existence of a phenomenon will require statistical analysis, but could interactive techniques assist analysts in the initial review of the data to find problems with the query, perceive patterns and formulate hypotheses about possible phenomena?

A second problem reported by our colleagues is that even highly trained medical professionals can forget the implicit but uncertain duration of medical events when interpreting temporal events, making the review process error-prone. To reuse the asthma example, researchers might be looking for patients who have been given steroids for their asthma condition. Because clinical data does not encode what condition a drug is prescribed for, users have to rely the data at hand to estimate why the drug was prescribed. Steroids are prescribed for asthma but also for rheumatoid arthritis or other long lasting medical conditions. Even though highly trained analysts should be well aware that rheumatism is a long-lasting condition, they are still susceptible to interpreting a rheumatism diagnosis, which appears as a point event, as a short-term condition, and make the wrong decision. Would showing the interval of validity of a diagnosis be helpful to remind users of the likely duration of the condition?

We propose Lifelines2, a prototype to visually explore multiple records of categorical temporal data. By allowing the alignment of data on sentinel events and showing intervals of validity (IV), we believe that these techniques can reduce unnecessary interaction and make visual review of temporal categorical data more effective.

RELATED WORK

Using timelines to present the temporal data is an obvious and increasingly common strategy. TimeSearcher [9] lets users specify patterns of interest to select and filter through thousands of temporal records of numerical data. Novel direct-manipulation widgets such as TimeBox and Angular Query allow users to capture the desired pattern without having to type. TimeSearcher2 [1] extends TimeSearcher by allowing multiple variables, and iterative flexible pattern searching. VizTree [12] enables users to find repeated patterns and anomalies in a large time series, such as electrocardiograms. The continuous time series is discretized into labeled bins. Using a tree, where each path corresponds to a pattern of labels, and the thickness of the path corresponds to the pattern's prevalence, patterns, anomalies, and motifs are accentuated. These approaches focus on patterns in numerical time series, while we focus on patterns in categorical data and their relationships across multiple records.

There has been a number of published visualization work on single patient record. These approaches focus on presenting raw medical readings on patients. Powsner and Tufte [18] integrate and display medical readings, each in a small time chart, making a compact graphical view that can include test results, X-ray images, and more. Bade et al. discuss a system that visualizes both quantitative data and ordinal data for medical readings of patients [2]. Quantitative data are discretized to better bring viewers' attention to the changes and extremes values of the readings. Concerned with missing values, Bade et al. allows intervals of validity for each reading point, so a 'best estimated value' is visually presented to the viewer. We extend their approach by allowing the intervals to start before a recorded data point to indicate that a condition may have existed prior to the reporting. While it is important to keep track of raw numbers, physicians often reason about a patient's condition on a conceptual level. Temporal abstraction techniques have been applied to medical record visualization [21, 17]. The focus is to allow users to build abstractions from many different readings, and make decisions about them on higher levels.

Aside from numerical data, there is also a rich literature on visualizing categorical data on timelines. Demographers have long used Lexis diagrams [12] and its variations [19] to visualize irreversible events in multiple life histories and to facilitate relative comparisons using a fixed temporal alignment. Lifelines [15,16] presents personal history record data organized in expandable facets and allows both point event and interval event representations. However, aside from panning, semantic zooming and text filters, there are very few ways to manipulate the data. Partners Healthcare currently uses Lifelines to display query results of large EHR databases [14].

PatternFinder [6] presents a form-based query interface for specifying temporal queries. The forms are very expressive, and give users extensive control in filtering. However, the queries are also complex to specify, and present a steep learning curve for new users. PatternFinder introduces a ball-and-chain visualization to display the complex set of matching results.

Research on interactive alignment of temporal data as a means of data manipulation is scarce. But there are approaches that align data by temporal periodicity. Aside from calendars, researchers have used spirals to visualize periodic data [4,22]. Hewagamage et al. propose a way to visualize events on spiral timeline in both 2D and 3D space, where events are represented as icons on the timeline [8]. However, the effectiveness of these

techniques rely on two factors: the periodicity of the data, and that an appropriate periodicity is chosen. These alignments are not interactive or data-driven. Experiscope [7] allows users to specify time points to align records of experimental results. The alignment serves similar purpose as our alignment – facilitating discovery of patterns – but we note that our approach does not require manual specification of alignment points in every record.

In the final stages of genome sequencing, researchers perform manual inspection to fix errors resulting from automated assembly. Hawkeye [20] facilitates this task by aligning matched DNA sequences and highlight problematic assemblies. A characteristic in genome assembly datasets is that every partial sequence is related (each is a part of the same DNA). Individual medical histories are, for most part, independent. DNA alignment supports finding the differences in different matches, while our approach focuses on finding similarities across histories. Another difference is that DNA sequences are discrete, on a uniform scale, and have only 4 nucleobases on the sequence. In contrast, temporal categorical data is far less constrained.

DESCRIPTION OF THE INTERFACE

Lifelines2 is an extension of Lifelines [15,16]. Lifelines was designed to summarize the entirety of a single personal history record (e.g. a medical record). In contrast, Lifelines2 displays selected subsets of the records from multiple patients. The output of a query (e.g. Find all patients who had a diagnosis of asthma and pneumonia-or-influenza) becomes the input data of Lifelines2. It is a Java application, utilizing the Piccolo 2D graphics framework [3]. We use real clinical data that had been de-identified to maintain privacy, and generated by the i2b2 query system under development at Harvard Medical School and Partners Health Care [14].

Each record is vertically stacked on alternating background color and identified by its ID on the left (Figure 1). Asthma and pneumonia diagnosis events appear as colored triangle icons on the timeline. By default all records are presented using the same absolute time scale (with the corresponding years or month labels displayed at the top) and the display is fitted so that the entire date range fits in the screen. As in Lifelines, zooming on the horizontal axis and panning is possible. Tool tips provide details, and records can be closed (one by one or all at a time) into a compact silhouette using smaller icons and less space. Left click onto the visualization centers and zooms in. Right click zooms out. Any click onto the record ID area resets the display to the initial fitted overview.

On the right side a control panel provides access to align, rank, and filter the display. Menus are data-driven. A user can choose any event category to align all the records. For example, Figure 2 shows setting the "*1st* pneumonia-or-influenza" as the sentinel event, and all records are aligned on a vertical line by that event. The time scale becomes relative and labels such as " +1 month", " -1 month" and so on. By default the 1st occurrence is used, but a menu allows users to switch to the next occurrences. Records that do not contain at least *n* occurrences of that event are filtered out of the display.

The records are listed in alphabetical order by default but users can rank records by the number of occurrences of a event category. In Figure 2 the records are ranked by the number of asthma events, bringing to the top the more severe cases. Users can also filter by the number of occurrences of events (e.g. removing records that contain only one pneumonia event). Users can also filter out records that do not contain a specified sequence of events (e.g. asthma followed by pneumonia). Finally the legend area can also be used to turn on and off certain types of events from the display to focus on a subset of event types.

With alignment we believe that Lifelines2 provides a simple yet effective mean of quickly exploring the data to look for potential temporal patterns across multiple records. When aligned, relative time spans can be compared easily, and one single zoom allows users to see the details around all sentinel events in view simultaneously. Overall the need to zoom and pan is greatly reduced, as is the need to keep in memory the scale of time ranges from record to record being compared. Ranking and filtering complement alignment by reordering or narrowing the set of records interactively to suit a user's changing focus. We affectionately call alignment, ranking, and filtering the ARF framework.

Existing applications have exploited alignment as a way to rearrange temporal or sequential data to reveal previously unknown patterns. However, these alignments are non-interactive [12,19], non-data-driven [4,22], or require manual specification of objects or time points to be aligned [7]. Our approach is solely based on data and is automatic in the sense that users do not specify alignment points for each record. Consequently it allows users to easily realign as their focus changes. The benefit of alignment in general may seem clear from its wide adoption in other applications. However, there is no previous work quantifying that benefit -- which we address in our evaluation -- nor study of its use in medical applications.

To test the idea of displaying intervals of validity, we built a control panel to allow us to specify a range before and after each event type. Intervals of validity are then displayed visually as a thin line extending from the point event in both temporal directions. The goal is to provide a visual reminder of the possible duration of the state or diagnosis represented by the point

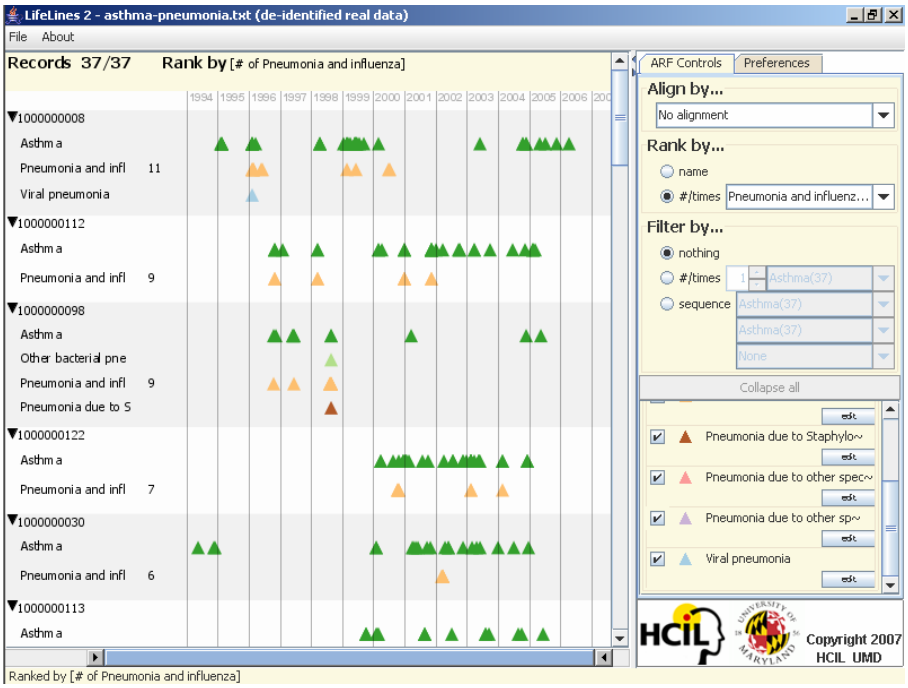

Figure 1. The interface without alignment. Each triangle represents an event. Note the data is presented chronologically, and the records are ranked by the number of *Pneumonia and influenza* events. It is easy to see the co-occurrence of *Pneumonia and influenza* and *Asthma*. However, it is not clear in patients' first *Pneumonia and influenza*, whether *Asthma* occurred before or after. Users are forced to zoom in to each first occurrence of *Pneumonia and influenza* for details, but each zoom can only reveal the details around a particular *Pneumonia and influenza* event.

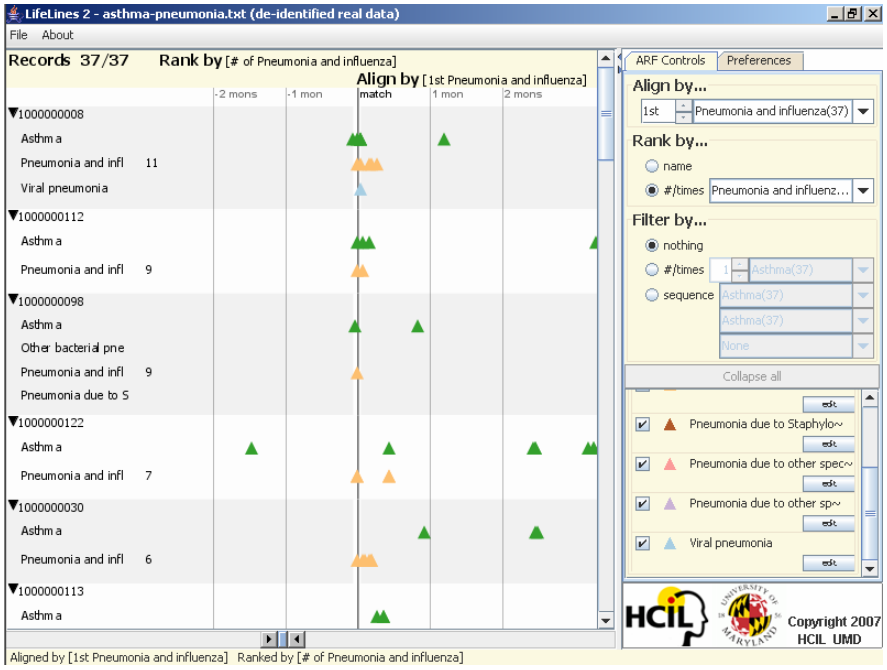

Figure 2. This figure shows the same dataset as in Figure 1. However, all patient records are aligned by the 1st *Pneumonia and influenza*. Note the relative time scale on the top. A single zoom had been applied to the alignment line. It is easily verifiable that the first 3 patients were diagnosed with asthma within a month prior or at the same time their pneumonia was diagnosed, while the other 2 patients in view were not.

event. Users no longer need to remember and estimate the duration of each event category, making it easier to spot events that occur concurrently (or in the case of clinical data events that "might" be occurring concurrently). Figure 3 shows sample data with and without the interval of validity. In this example the interval values are suggested by our physician colleagues and specified manually by us. Ultimately the length of the intervals would be specified by a trusted, authoritative data provider, or possibly computed based on other factors (e.g. age of the patient). Our controlled approach allows us to study how the intervals are interpreted by users and measure if the intervals improve performance at least in the simplest tasks.

EVALUATIONS

We conducted two separate user studies. In the first study we aimed to quantify the benefits of alignment and intervals of validity using a controlled experiment. Our goal in this experiment was also to observe what

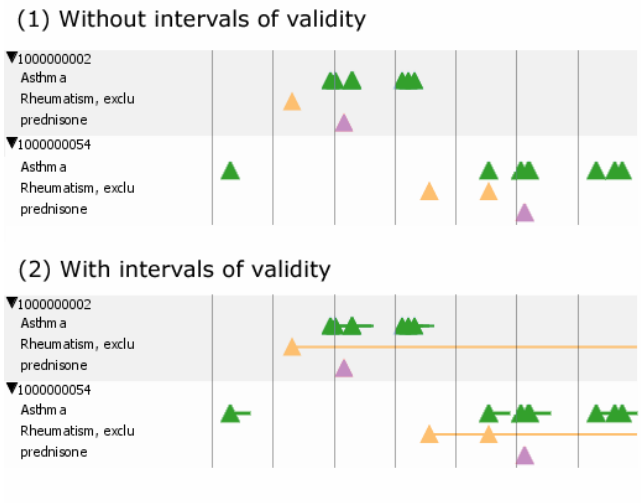

Figure 3. The top portion shows two partial records without intervals of validity. Because prednisone (a steroid) prescriptions coincided temporally with asthma diagnoses and with no other events, users may mistakenly conclude that prednisone was given for asthma. The bottom portion shows the same two records with intervals of validity. Rheumatism's lasting interval allows users to visually confirm that there may be more than one reason why predisone was prescribed.

strategies users chose, and what problems they would encounter. Because medical professionals have very little availability, and are hard to recruit for a user study, this 1st study used data and experts from another domain. We used synthetic data based on graduate school academic events and recruited graduate students, faculty, and staff who are familiar with graduate academic life. We designed the tasks similar to the tasks medical researchers would perform, and verified that the tasks were representative with medical experts. In the second user study, we interviewed medical professionals and used real, but de-identified medical data. The goal of the second study was to obtain domain experts' comments, suggestions, reflections on the uncertainty aspects of the data, and some preliminary quantitative data as well.

Controlled Experiment - Procedure

The data consisted of scholastic records for a set of students. Data were point events such as submitting a paper, a software release, dates of a dissertation proposal or defense, signing up for a class, or submitting a job application. There were 20 participants in this study.

The experiment was further divided into two independent parts. The first part evaluated the benefits of alignment. The second part evaluated the benefits of showing the intervals of validity. Both parts of the experiment follow a repeated measures design. Each part had 2 interface variations, and each participant performed a set of tasks once for each variation. We recorded the time to complete each task and errors, if any.

In Part 1, participants were asked to perform tasks regarding events around sentinel events, using 2 interface variations of Lifelines2. One variation had alignment as an operator (ARF), while another did not (RF). We gave a short demonstration of the interfaces, answered questions, and let the participants familiarize themselves with the interface (for a total 15 minutes of training). We then asked them to perform specific tasks. Because each participant needs to decide on how to best approach a task, we allowed each participant to read each task description and formulate a plan of attack before loading the task data and starting the timer. This ensured that the time we recorded was the time of task completion, and did not include plan formulation or task comprehension. We recorded the strategies the participants used. Every participant was asked to perform the set of tasks below, once with the alignment feature available (i.e. with ARF), once without (i.e RF). The order in which the two variations of the interface were presented was counter-balanced to mitigate learning effects. The tasks and their design rationale for this part of the experiment are discussed below.

- Task 1: How many students submitted a paper within 1 month after proposal? (5 records)
- Task 2: How many students submitted a paper within 1 month after proposal? (20 records)
- Task 3: How many students published at least 3 papers between proposal and defense?
- Task 4: What occurred most often within a month of a student's 1st paper submission?

Task 1 and 2 are similar to tasks where a researcher must study the relationship between a sentinel event and another temporally related event category. We note that the sentinel event was deliberately made clear in these two task descriptions. While the task descriptions for the

two tasks were the same, the data in Task 1 was much simpler than that in Task 2. Task 1 included only 5 records that fit on the screen, so users could find the correct answer without having to interact at all with the display. Task 2 had more records (20) with an expanded time range so users needed to zoom, pan, and scroll significantly to answer the question correctly. We first hypothesized that alignment would reduce the time it took users to perform these tasks even when all the data was available on one screen. We further hypothesized that when more interaction was required, the benefits of alignment would be much greater.

Task 3 is conceptually more complex and there was no clear way to manipulate the data using alignment, ranking, and filtering to find the answers. This task simulated the process of temporal pattern confirmation researchers might perform (such as confirming a hypothesis about certain patterns of events). While this task required participants to focus on temporal range comparisons relative to a sentinel event (proposal), it did not require detailed inspections that incur extensive interactions. Therefore while we expected to observe alignment's benefits, we also expected them to be less pronounced.

Finally, Task 4 simulated how a researcher would go about discovering new patterns around sentinel events (in this case, a simple temporally-constrained co-occurrence relationship). We hypothesized that the benefits of alignment would be significant, since both intense interaction and relative comparisons were required in this task.

In the second part of the experiment, we used only the interface variation with alignment (ARF), but we varied how events were represented. In one condition (IV) the lines of the interval of validity were visible; in the other condition (no IV) they were not. Participants performed the set of tasks below, once for each variation. The order in which the interface variations were shown to the users was also counter-balanced. Finally, the participants were asked to fill out a subjective satisfaction questionnaire for each part of the experiment. The two tasks for this part are listed below.

- Task 5: Assuming a class lasts 3 months, how many students proposed while they were taking a class?
- Task 6: Assuming a class lasts 3 months, and it takes 2 months to prepare for proposal, how many students were preparing for proposal while taking a class?

In Task 5, participants were to find proposals that occur within 3 months after a class-signup event, and only the interval of the class-signup events was of concern. In Task 6, however, both the intervals of class-signup and event and the proposal event came into play.

We hypothesized that when intervals of validity were shown visually, participants could perform both tasks 5 and 6 more quickly and with lower error rate. In addition, because there were more intervals to keep track of, increasing cognitive load, the benefits of the intervals will be more dramatic in Task 6. Since we provided the duration of the events in the task description, this 1st pilot study did not address the hypothesis that the lines might remind users of the likely duration of the event. This was only addressed in the qualitative study described in a later section.

We used a different dataset for each task in each interface variation. There were 2 sets of datasets, one for each interface for Part 1. The same applied for Part 2. The data complexity was comparable between any corresponding tasks in each set. All experiments were conducted on an IBM laptop with the following specifications: 14 inch screen, 1.4 Ghz CPU, 1.2GB RAM, Windows XP Professional. Every participant used an optical mouse we provided.

Controlled Experiment - Results

We analyzed each task separately. We used a repeated measures one-way ANOVA to verify the significance of the time differences in task completion. We used Cochran-Mantel-Haenzsel general association statistic to test whether the error rates between two interfaces were different in both parts of the experiment ($p=0.05$). Finally, for numerical answers that participants gave, we used a one-way ANOVA to see if the difference in the size of the errors was significant. Figures 4, 5, and 6 show the results, and we discuss them here in detail.

Throughout the experiment, participants were observed to use the following effective strategy: first reduce the data via filtering and alignment (if available), and only manually inspect data that have potential. Ranking was not used very often.

In Task 1 and 2, when using ARF (i.e. with alignment), all except two participants chose to use alignment or alignment in conjunction with a filter. When alignment was not available (i.e. in the RF interface variation), most participants used sequence filter.

In Task 1, there was no statistical difference in term of completion time. In Task 2, the benefits of alignment were statistically significant. Participants were able to complete the task 65% faster with alignment ($p < 0.0001$). In addition, participants were less prone to make errors when using alignment ($p < 0.02$). When they did make an error, the size of error was significantly smaller than when alignment is not available ($p < 0.05$). These results supported our hypothesis that alignment reduces disruptive interaction and allows users to perform these tasks faster and more accurately when the data is complex (i.e. more than a few records).

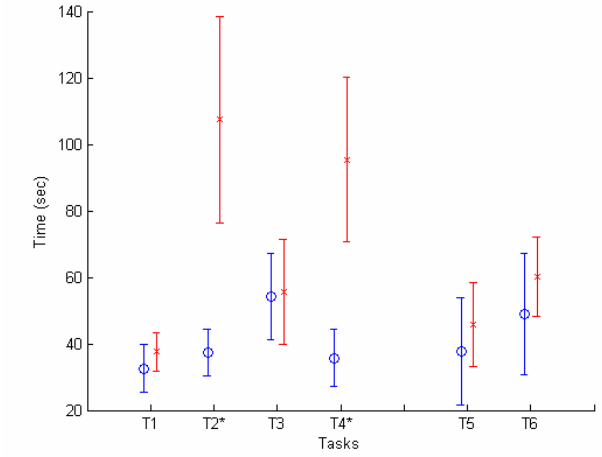

Figure 4. Task Completion Time. Each circle or x denotes the mean, and each vertical line denotes the standard deviation. Blue marks denote the results for the align-rank-filter (ARF) variation of the interface in Task 1 to 4 and interval of validity (IV) variation in Task 5 and 6. Although blue marks always have lower means, only in Task 2 and 4 were there statistical significance (denoted by an asterisk in task names). The significance was salient ($p < 0.0001$) in both cases.

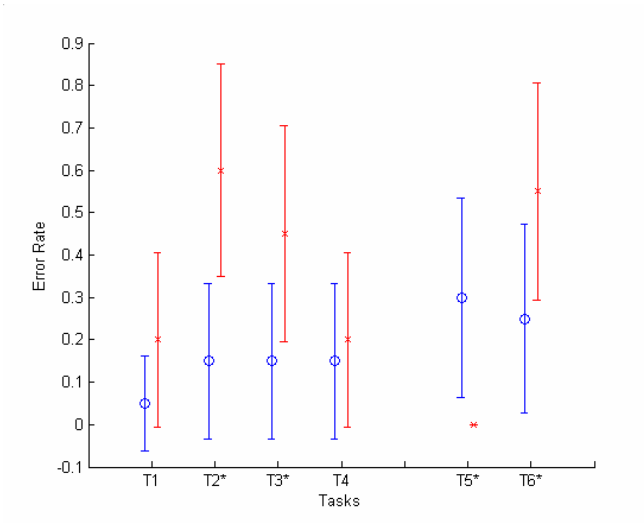

Figure 5. Error Rate by Tasks. Blue marks represent align-filter-rank (ARF) variation of the interface in Tasks 1 to 4 and interval of validity (IV) variations for Task 5 and 6. There was significance in error rate in Task 2 ($p < 0.01$), Task 3 ($p < 0.1$), Task 5($p < 0.05$), and Task 6 ($p < 0.1$). We noted that in Task 5, users performed better when intervals of validity were *not* displayed.

As we expected, participants had very different strategies in completing Task 3 because the fastest strategy was not as clear as in Task 1 or 2. When alignment was an option, over two-thirds of the participants aligned by proposal and used a type of filter. Most popular filters were "at least 3 paper submissions", and the sequence "proposal, paper, defense". No participant tried to align or filter solely by defense, although it was the most effective technique in this question (students must have a proposal before they can defend). This reflected that it takes time to learn to make the best use of any new feature such as alignment. When alignment was not available, sequence filter was the favorite among the participants. The sequences "proposal, paper, defense" and "proposal, defense" were the most popular.

In Task 3, the difference in completion speed was negligible (2.9%). Our participants tended to make an error less often when alignment was available ($p < 0.1$), however, there were no significant differences in error size. This result did not support our hypotheses on completion time, though the hypothesis regarding error rate was somewhat supported.

The performance difference in task completion again was startling in Task 4. Using alignment, our participants were able to complete 62% faster ($p < 0.0001$) with no significant difference in error rates. All but 1 user used alignment when alignment was available. When alignment was not an option, most users opted to use "paper submission" as a filter but seven participants used neither filter nor ranking.

In the second part of the experiment Task 5 and Task 6 were designed to measure the possible benefits of intervals of validity in terms of the time it takes to visually find potential overlaps. Our hypothesis that intervals of validity help users perform faster and more accurately was not supported. We found that there were no differences in completion time. There were significant but conflicting differences in error rates. Users were less likely to make a mistake when intervals of validity were not displayed in Task 5 ($p<0.05$). In Task 6, users were less likely to make a mistake (by a margin of 30%, $p<0.1$) when intervals of validity were displayed.

Task 5's result was puzzling, but one participant offered one possible explanation in her comments. She noted that she relied on visualization more when intervals of validity were displayed, and was less inclined to zoom-in to verify her answers, especially when the task is as simple as Task 5.

An interesting observation was that every participant used the alignment function in both Task 5 and Task 6, regardless of whether they had been particularly successful with alignment in the first part of the study. We hypothesized that participants recognized that Tasks 5 and 6 were relative comparison tasks, and understood (presumably from previous experiences with it) that alignment would get them to the answer the fastest.

In the questionnaire users rated alignment very positively. On a scale of 1 to 9, users agreed that alignment was helpful in Tasks 1 to 4 with a mean of 8.3. They also agreed that they were able to perform faster in Tasks 1 to 4 when alignment was available with a mean rating of

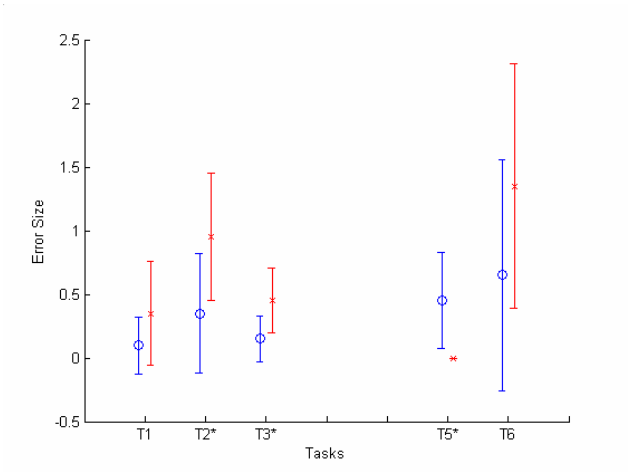

Figure 6. Error Size by Task. Responses for Task 4 were not numeric, so error size was not applicable. Blue marks denote the ARF variation of the interface in Tasks 1 to 3 and IV variations in Tasks 5 and 6. Error sizes were significantly smaller for ARF variation in Task 2 ($p < 0.05$) and Task 3 ($p < 0.1$). Error sizes were significantly smaller in Task 5 ($p < 0.05$), in favor of the no IV variation.

7.95. One participant commented that "scrolling was a big distraction, and alignment helped eliminating a lot of it". Another participant commented that when using the interface where alignment was not an option, he noticed that he was mentally doing the alignment himself, suggesting that alignment is a natural operator when people need to reason about time. While intervals of validity were still rated as helpful in Tasks 5 and 6 (7.85), the overall perception was that they were not as critical as alignment. Our participants believed that they could perform tasks 5 and 6 more quickly with intervals of validity with a rating of 7.55. One participant did comment that the intervals of validity allowed him to get an idea how long the durations are, "especially if I forgot."

Domain Expert Qualitative Evaluation - Procedure

In this initial study we interviewed 4 medical professionals -- one registered nurse, one physician, and two faculty members in a nursing school, none had a connection with the development of this interface. Three have had experience with medical informatics systems and have been involved in medical research. We showed them our application using de-identified data from our colleagues at Harvard. We asked them to talk out loud while using the application, without any training. We recorded how they interpreted the visual display and their impressions of the interface, the ARF framework, intervals of validity, usability issues, and suggestions to future refinements.

We first loaded a dataset corresponding to the problem described in the Motivation section regarding the uncertain nature of the data, and how it was to be addressed by the intervals of validity. It contained 11 patients who had asthma, and were prescribed some type of steroid. Some of the patients had other conditions (e.g. rheumatoid arthritis, pneumonia) that might also require steroids treatments. We gave the only introduction that Lifelines2 is used to view results from a search in patient databases. We asked our interviewees to comment on what they saw, and how they interpreted the display.

Once they had reviewed the data, we asked them to imagine that they were selecting patients for a study, and we only wanted patients who were given steroids for their asthma condition. Those who were given steroids for other reasons should not be included. Of course, there was no way to know the actual reason why a patient was prescribed steroids, but we wanted them to select likely ones, and filter out the ones that seem unsuitable to trigger some reasoning about the uncertainty of the data. We then asked them to go thru the same exercise again with the same data, but with the intervals of validity visible (Figure 3 shows the differences in display). We used interval values provided by our colleague at Harvard medical school. Rheumatoid arthritis and rheumatisms were given the interval of infinity. Asthma had 4 months, and pneumonia had about 2 weeks. We asked our experts how they would interpret the intervals. We also looked at how their answers differed in the two sessions, and asked how the intervals made them change their mind.

We then loaded the data containing patients with asthma and pneumonia. We asked our medical experts to look at the data and see if they could find trends in the dataset. In particular, whether there were more patients who first contracted pneumonia before an asthma event or there were more of those who first had asthma, then pneumonia. We showed them how to use alignment, ranking, and filtering.

Then, using another dataset containing 45 patients with various heart problems, we asked our experts to use the interface on their own and to discuss what might be the most commonly co-occurring event to an acute myocardial infarction in this set of patients. This time around, we let them go about the task without much guidance or interruption. We aimed to observe whether ARF made sense to them, and how they would utilize its functionality on their own.

Domain Expert Qualitative Evaluation - Results

Three out of four participants had no problem interpreting the visualization. They were able to immediately figure out the timeline, each patient's record, and the temporal event data shown as triangles. One participant did not immediately grasp that individual patient records were displayed. She had interpreted the triangles to indicate sets of patients at the beginning, but soon after realized on her own what was going on. Two participants commented on how helpful the color encoding was.

Looking at the asthma-steroid data, all of them talked about the asthma triangles as flares, not merely diagnoses. When asked to find patients who were given steroids for their asthma condition, most of them employed the strategy we had anticipated: find each steroid prescription event, and see what the closest event prior to that was. If the closest event was asthma, and it was within a few months, then they are likely to think the steroid was prescribed for asthma. Two participants took into account the frequency of how correlated the steroid prescription event is to other events. When no intervals of validity were displayed, none of the participants paid much attention to the possibility of a long lasting rheumatism condition. When the intervals of validity were displayed, they all interpreted them as durations. However, two participants interpreted the intervals as certain or known duration, and not as an uncertain, possible duration.

The important result was that the displayed intervals affected how our participants viewed and interpreted the data. When they were asked to find patients who were given steroids for asthma with intervals of validity, the lines reminded them the durations of each event, and some of their answers changed from the no interval version, although the direction of change did not always conform to what we expected. We had expected them have picked fewer patients for the clinical trial because they might have forgotten or ignored an existing rheumatoid arthritis. However, two of the participants picked more patients because some unexplained steroid prescription event is now explained by an asthma that occurred just 3 months before. This was now easy to see with the 4 months duration for each asthma event. Two participants did not like that fact that someone else could assert intervals because they might not trust that person. They did agree that if they had added the intervals themselves, tailored to the questions they have in mind, then they would have no objections. One participant commented that "intervals of validity were good reminders of uncertainty, and using the range to perform overlapping tasks is much simpler." He also commented that long intervals draw a lot of attention, and whether such attention is good might depend on the situation.

Our participants were quick to grasp the ARF framework of align, rank and filter. One participant figured out how to use them on her own without any introduction or demonstration of the functionalities. Using the asthma-pneumonia data, all participants were able to quickly align or filter to see if there were overall patterns in those patients. They did comment that alignment and ranking allowed them to figure out the data quickly. They also liked that once the data is aligned by a patient's 1st pneumonia, zooming in and scrolling down were all it took to get through all the patients efficiently. One participant commented that grouping records by whether a second sentinel event occurred before or after the 1st one might be a good addition. Ranking using temporal relationships to aligned sentinel events was also mentioned as being potentially useful.

When we loaded the set of patients with heart diseases and asked our participants to discover patterns of events with regard to acute myocardial infarction (AMI), 2 out 4 quickly approached the question with alignment on the correct event. They were able to see that coronary atherosclerosis co-occurred with AMI most often in this dataset. One participant, using her medical knowledge, found another association pattern between coronary atherosclerosis and hypertensive disease events in this data.

Two participants quickly offered examples where the application could be useful in their own research and became very enthusiastic. They both were involved in working with clinical data (though independent with each other), and were impressed on how effortless it was to identify patterns using the ARF framework. One example was to use this approach in a study of patterns of events that might suggest medical misdiagnosis in an emergency room.

In addition to this pilot study the prototype was demonstrated to our sponsors and partners and received unusually encouraging feedback. During a presentation at the National Institute of Health a respected medical researcher at the University of Chicago commented, "[it will] change the entire paradigm in medicine".

CONCLUSIONS AND FUTURE WORK

Lifelines2 provides primitive operations to align, rank and filter the results of queries. Displays of patient histories aligned on sentinel events enable medical researchers to spot precursor, co-occurring, and aftereffect events. A controlled study with 20 participants demonstrated the benefits of providing alignment for larger sets of records (with up to a 60% improvement). We believe our ARF framework can be expanded to allow users to interactively, incrementally, and systematically discover previous unseen patterns. A pilot qualitative study with four medical professionals suggested that the interface can be learned quickly and addresses the need to rapidly review results and spot patterns of interest. Our qualitative study also revealed that there is a lot more to be done in terms of providing a visual representation that adequately represents the "messiness" of clinical data.

Clinical data tend to be messy with aspects that become only obvious when the data is visualized. The same heart attack might be recorded three times in three days (by the emergency room physician, a cardiologist, and a clerk from the billing office) and it can be hard to differentiate it from 3 separate events. Even if medical event information is carefully recorded at the time of the doctor visit or during a hospitalization, the time stamp is usually inaccurate by nature. When a new patient comes into a doctor's office complaining of shortness of breath and a

diagnosis of asthma is recorded, the time stamp represents the time the diagnosis was recorded in the system, not when an asthma attack may have occurred. When the condition first occurred remains unclear. Furthermore, since asthma is a condition that usually lasts for while, a diagnosis implies that the asthma condition persisted for the next few months. Representing these events as points in a time line is simple but it seems insufficient. Our attempt at representing them as lines may also be too rudimentary. Our observations confirmed that the lines helped users remember about long durations of events, but new problems appeared when users interpreted the line as an actual confirmed duration. Future research will need to investigate alternate representations for uncertain data and to study how they are interpreted.

ACKNOWLEGEMENT

This project is supported in part by the Washington Hospital Center and Harvard - Partners HealthCare. This work is also supported in part by grants from Fujitsu, Lockheed Martin, NTT Corp., Kevric Corp., SAIC, the National Science Foundation, the National Geospatial Intelligence Agency, DARPA, US Army Research Laboratory, and NIST.

REFERENCES

1. Aris, A., Shneiderman, B., Plaisant, C., Shmueli, G., Jank, W., Representing unevenly-spaced time series data for visualization and interactive exploration. *Pro. INTERACT 2005*, LNCS 3585, 835-846.

2. Bade R., Schelchtweg S., Miksch S., Connecting time-oriented data and information to a coherent interactive visualization. *Proc. CHI 2004*, ACM Press (2004).

3. Bederson, B. B., Grosjean, J., Meyer, J., Toolkit design for interactive structured graphics. *IEEE Tran. On Software Engineering*, 30(8), pp. 535-546.

4. Carlis, J., Konstan, J., Interactive visualization of serial periodic data. *Proc. UIST 1998*.

5. Chung, W., Chen, H., Chaboya, L. G., O'Toole, C., Atabakhsh, H., Evaluating event visualization: a usability study of COPLINK spatio-temporal visualizer. *IJHCS*, 62, 1, 127-157, 2005.

6. Fails, J., Karlson, A., Shahamat, L., Shneiderman, B., A visual interface for multivariate temporal data: Finding patterns of events over time. *Proc. VAST 2006*.

7. François Guimbretière, Morgan Dixon and Ken Hinckley. ExperiScope: An analysis tool for interaction data. *Proc. CHI 2007*, pp 1333 – 1342.

8. Hewagamage, K.P., Hirakawa, M., Ichikawa, T., Interactive visualization of spatiotemporal patterns using spirals on a geographical map. *Proc. VL 1999*, 296-303, 1999.

9. Hochheiser, H., Shneiderman, B., Dynamic query tools for time series data sets, timebox widgets for interactive exploration. *Information Visualization 3*, 1 (2004), 1-18.

10. i2 Software. http://www.i2inc.com/Solutions/Counterterrorism/

11. Lam, H., Russell, D. M., Tang, D.,Munzner, T., Session viewer: supporting visual exploratory analysis of web session logs. *Proc. VAST*, 2007

12. Lexis, W., *Einleitung in die theorie der bevölkerungsstatistik*, Strassburg: Trübner, 1875.

13. Lin, J., Keogh E., Lonardi, S., Lankford, J., Nystrom, D., Viztree: Visually mining and monitoring massive time series. *Proc. ACM SIGKDD*, 2004.

14. Murphy, S., Mendis, M., Hackett, K., Kuttan, R., Pan, W., Phillips, L., Gainer, V., Berkowicz, D., Glaser, J., Kohane, I., Chueh, H., Architecture of the open-source clinical research chart from informatics for integrating biology and the bedside. *Proc. AMIA*, 2007

15. Plaisant, C., Milash, B., Rose, A., Widoff, S., Shneiderman, B., Lifelines: visualizing personal histories. *Proc. CHI 1996*.

16. Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D., Shneiderman, B., LifeLines: Using visualization to enhance navigation and analysis of patient records. *Proc. AMIA Fall Symposium*, 1998, 76-80.

17. Post, A. R., Harrison, J. H.. Protempa: A method for specifying and identifying temporal sequences in retrospective data for patient selection. *JAMIA*, 2007.

18. P Powsner S., Tufte E., Graphical summary of patient status. *The Lancet*, 344 (August 6, 1994), 386-389.

19. Pressat, R., *L'analyse démographique. méthodes, résultats, applications.* Paris: Presses Universitaires de France, 1961.

20. Schatz, M.C., Phillippy, A.M., Shneiderman, B., Salzberg, S.L., Hawkeye: an interactive visual analytics tool for genome assemblies. *Genome Biology*, 2007

21. Shahar, Y., Cheng, C., Intelligent visualization and exploration of time-oriented clinical data. *Proc. HICSS 1999*.

22. Weber, M., Alexa, M., Müller, W., Visualizing time-series on spirals. *Proc. INFOVIS 2001*, 7-14, 2001.