High-Volume Hypothesis Testing: Systematic Exploration of Event Sequence Comparisons

SANA MALIK, BEN SHNEIDERMAN, FAN DU, CATHERINE PLAISANT,
and MARGRET BJARNADOTTIR, University of Maryland, College Park

Cohort comparison studies have traditionally been hypothesis driven and conducted in carefully controlled environments (such as clinical trials). Given two groups of event sequence data, researchers test a single hypothesis (e.g., does the group taking Medication A exhibit more deaths than the group taking Medication B?). Recently, however, researchers have been moving toward more exploratory methods of retrospective analysis with existing data. In this article, we begin by showing that the task of cohort comparison is specific enough to support automatic computation against a bounded set of potential questions and objectives, a method that we refer to as High-Volume Hypothesis Testing (HVHT). From this starting point, we demonstrate that the diversity of these objectives, both across and within different domains, as well as the inherent complexities of real-world datasets, still requires human involvement to determine meaningful insights. We explore how visualization and interaction better support the task of exploratory data analysis and the understanding of HVHT results (how significant they are, why they are meaningful, and whether the entire dataset has been exhaustively explored). Through interviews and case studies with domain experts, we iteratively design and implement visualization and interaction techniques in a visual analytics tool, CoCo. As a result of our evaluation, we propose six design guidelines for enabling users to explore large result sets of HVHT systematically and flexibly in order to glean meaningful insights more quickly. Finally, we illustrate the utility of this method with three case studies in the medical domain.

CCS Concepts: • **Human-centered computing** → **Visualization**; *Visualization systems and tools;*

Additional Key Words and Phrases: Cohort comparison, event sequences, visual analytics

ACM Reference Format:
Sana Malik, Ben Shneiderman, Fan Du, Catherine Plaisant, and Margret Bjarnadottir. 2016. High-volume hypothesis testing: Systematic exploration of event sequence comparisons. ACM Trans. Interact. Intell. Syst. 6, 1, Article 9 (March 2016), 23 pages.
DOI: http://dx.doi.org/10.1145/2890478

1. INTRODUCTION

Sequences of timestamped events are currently being generated across nearly every domain of data analytics. Consider a typical e-commerce site tracking each of its users through a series of search results and product pages until a purchase is made. Or consider a database of electronic health records containing the symptoms, medications, and outcomes of each patient who is treated. Every day, this data type is reviewed by

The reviewing of this article was managed by the associate editors of the special issue on Highlights of IUI 2015, Shimei Pan and Giuseppe Carenini.
This work is supported by Adobe and the University of Maryland/Mpowering the State through the Center for Health-related Informatics and Bioimaging (CHIB).
Authors' addresses: S. Malik, B. Shneiderman, C. Plaisant, and F. Du, Human-Computer Interaction Lab, University of Maryland, College Park, MD 20742; emails: {maliks, ben, plaisant, fan}@cs.umd.edu; M. Bjarnadottir, Robert H. Smith School of Business, University of Maryland, College Park, MD 20742; email: margret@rhsmith.umd.edu.

DOI: http://dx.doi.org/10.1145/2890478

Fig. 1. Two datasets, each containing about 1,000 patients as they are transferred throughout a hospital, are being compared using CoCo: patients who lived versus patients who died (demo dataset; no real data). On the left, a legend pairs each event type with a color marker. Below, high-level statistics about each dataset are shown in a table. The center panel displays a compact view of the results of high-volume hypothesis testing, ranked by significance. Within the list, details-on-demand for a selected hypothesis (comparing the average timing between the blue and red event) is expanded in the gray panel to provide more details and context for the results. To the right of the result list is a minimap of the results, which provides an overview of the users' progression through the result set, showing how many results have been reviewed (gray) and how many are left (red). A set of control panels (bottom left and right panel) allows users to sort and filter the results by event sequence length, event types, sample size, significance, or metric type.

humans who apply statistical tests, hoping to learn everything they can about how these processes work, why they break, and how they can be improved upon.

Human eyes and statistical tests, however, reveal very different things. Statistical tests show metrics, uncertainty, and statistical significance. Human eyes see context, confirm what they already know, and discover patterns that are unexpected.

Visualization tools strive to capitalize on these latter, human strengths. For example, the EventFlow visualization tool (http://hcil.cs.umd.edu/EventFlow) [Monroe et al. 2013a] supports exploratory, visual analyses over large datasets of temporal event sequences. This support for open-ended exploration, however, comes at a cost. The more that a visual analytics tool is designed around open-ended questions and flexible data exploration, the less it is able to effectively integrate automated, statistical analysis. Automated statistics can provide answers, but only when the questions are known.

The opportunity to combine these two approaches lies in the middle ground. By all accounts, the goal of open-ended questions is to generate more concrete questions. As these questions come into focus, so too does the ability to automatically generate the answers. This article introduces CoCo (for "Cohort Comparison," Figure 1), a visual analytics tool that is designed to capitalize on one such scenario.

Consider again the information that is tracked on an e-commerce site. From a business perspective, the users of the site fall into one of two groups: people who bought something and people who did not. If the goal is to convert more of the latter into the former, it is critical to understand how these two groups, or cohorts, are different. Did one group look at more product pages? Or spend more time on the site? Or have some clear demographic identifier such as gender, race, or age? Similar questions arise in the medical domain as well. Which patients responded well to an experimental

medication? How did their treatment patterns differ from the patients who received the standard treatment?

Although comparing two groups of data is a common task, with temporal event sequence data in particular, the task of running many statistical tests becomes complex because of the variety of ways the cohorts, sequences (entire records), subsequences (a subset of events in a record), and events can differ. In addition to the structure of the event sequences (e.g., order, co-occurrences, or frequencies of events), the attributes about the events and records (e.g., gender of a patient) and the timestamps themselves (e.g., an event's duration) can be distinguishing features between the cohorts. For this reason, running statistical tests to cover all these cases and determining which results are significant become cumbersome. Based on 3 years of case studies, we present a taxonomy of metrics for comparing cohorts of event sequences.

Current tools for cohort comparison of temporal event data (Section 2) emphasize one of two strategies: (1) purely visual comparisons between groups, with no integrated statistics, or (2) purely statistical comparisons over one or more features of the dataset. By contrast, CoCo is designed to provide a more balanced integration of both human-driven and automated strategies.

We begin by showing that the task of cohort comparison is specific enough to support automatic computation against a bounded set of potential questions and objectives, a method we refer to as High-Volume Hypothesis Testing (HVHT). From this starting point, we demonstrate that the diversity of these objectives, both across and within different domains, as well as the inherent complexities of real-world datasets, still requires human involvement to determine meaningful insights. We explore how visualization and interaction better support the task of exploratory data analysis and understanding HVHT results (how significant they are, why they are meaningful, and whether the entire dataset has been exhaustively explored). Through interviews and case studies with domain experts, we iteratively design and implement visualization and interaction techniques in a visual analytics tool, CoCo. As a result of our evaluation, we propose six design guidelines for enabling users to systematically and flexibly explore large result sets of HVHT in order to glean meaningful insights more quickly. Lastly, we illustrate the utility of this method with three case studies in the medical domain.

The direct contributions of this article are as follows:

(1) A taxonomy of metrics for comparing groups of temporal event sequences
(2) Six design guidelines for providing systematic exploration of high-volume hypothesis testing (HVHT) results
(3) A visual analytics tool that implements these design guidelines for balanced integration of HVHT and user-guided analysis with an intelligent user interface
(4) Three case studies in the medical domain that illustrate the applicability and utility of the proposed visualization and interaction designs for the analysis and comparison of medication usage patterns

On a broader level, the goal of this article is to highlight the relationship between task specificity and the ideal balance between humans and statistical analysis, so that future efforts can better leverage the strengths of both approaches.

2. RELATED WORK

2.1. Event Sequence Visualization and Comparison

Gleicher et al. [2011] provide an extensive survey of visual comparison techniques classified into three categories and combinations thereof: juxtaposition, superposition, and explicit encoding. We use this characterization as a framework for exploring

designs for visualizing comparison results. Though many visualization tools have been designed for event sequence visualization [Monroe et al. 2013a; Stolper et al. 2014], there has been little research on visualizing event sequence comparison until recently. Zhao et al. [2015] design MatrixWave, a visualization designed to compare the flow of users in clickstream datasets. MatrixWave focuses on differences in the occurrence of immediate, pairwise steps in the event stream, whereas we generalize to differences in single events and sequences of any length, as well as differences dealing with time.

In addition to finding differences in datasets, event sequence comparison has been explored in the context of finding similarities. Vrotsou et al. [2014] introduce a set of event sequence similarity measures. They explore using visualization and interactive data mining to cluster similar groups of event sequences. While we focus on difference metrics, we believe our work can be extended to applicable similarity measures.

2.2. Statistics for Comparing Cohorts

In medical cohort studies, the most prevalent approach for comparison is survival analysis, where survival time is defined as the time from a given point to the occurrence of an event [Bewick et al. 2004]. The Kaplan-Meier method is often used to analyze the survival time of patients on different treatments and to compare their risks of death [Collett 2003; Dupont et al. 2004; Goel et al. 2010]. Based on the Kaplan-Meier estimate, the survival time of two groups of patients can be visualized and compared with survival curves, which plot the cumulative proportion surviving against the survival times [Bewick et al. 2004]. Also, the log-rank test is often used to statistically compare two survival curves by testing the null hypothesis (i.e., that there is no difference between the curves). Compared with survival analysis, the event sequences data used in our work is much more complicated and requires a more advanced analysis model.

Currently, tools that combine visualization and statistics for medical cohort analysis focus on single cohorts. CAVA [Zhang et al. 2014] is a visualization tool for interactively refining cohorts and performing statistics on a single group. Recently, Oracle published a visualization tool for cohort study [Oracle 2011]. Based on patients' clinical data, it supports interactive data exploration and provides statistics as well as visualization functionalities. These tools similarly focus on combining visualization with automated statistics and providing an interactive interface for selecting cohorts; however, both tools aim at grouping and identifying patient cohorts for further characterization, while our work focuses on comparing two existing cohorts based on their event histories.

2.3. Exploratory Hypothesis Testing

As event sequence datasets grow larger and larger, researchers are moving toward more exploratory methods for hypothesis generation and testing. The statistical implications of high-volume hypothesis testing (e.g., inevitable false positives) have been extensively researched [Benjamini and Yekutieli 2001; Shaffer 1995; Dunn 1961]. In CoCo, we allow users to apply the Bonferroni correction [Dunn 1961], which corrects for false positives by dividing the acceptable significance level by the number of tests performed.

Liu et al. [2011] explore the statistical and technical implications of automatically generating and testing many hypotheses. Similar to our work, they find that interactive techniques such as sorting and filtering are necessary for parsing these result sets, but their display is largely textual. We explore more visual methods for displaying both the hypothesis and results.

2.4. Temporal Data Mining

Automated hypothesis testing is closely related to big data mining. There are many established algorithms for frequent sequence mining [Mabroukeh and Ezeife 2010;

Han et al. 2007] and association rule (itemset) mining [Agrawal et al. 1993]. The majority of data mining techniques focus on mining sequences in a single dataset and not comparing across two datasets. While two data mining techniques can be used in tandem to facilitate similar comparisons (e.g., comparing frequent sequence results across two datasets), more specialized methods are needed to answer the question, "Which sequences occur significantly differently between these datasets?" Bay and Pazzani [2001] introduce contrast mining sets, an algorithm for detecting differences between groups based on record attributes, such as age, gender, or occupation. In addition to record attributes, we also look at differences in event sequences, based on both occurrence and timestamps.

Typical data mining algorithms are a blackbox, allowing little user involvement during the process. Recent work has been done on interactive sequence mining [Vrotsou and Nordman 2014; Lammarsch et al. 2014; Perer and Wang 2014; Federico et al. 2015], though these systems focus primarily on mining frequent patterns in a single dataset. Little work has been done on involving the user in mining differences between datasets.

3. BALANCING AUTOMATION WITH HUMAN INTERACTION

Purely statistical methods of comparison would benefit from user intervention. With the sheer number of metrics, it is time consuming to run every metric before exploration can begin, especially when not every metric may be required for analysis. Users with domain knowledge about the datasets would ideally be able to select from the metrics and easily eliminate unnecessary metrics. Further, questions asked during cohort comparison may vary based on how the cohorts were divided. If the cohorts were divided by outcome (e.g., patients who lived vs. patients who died), the sequence of events leading up to them becomes more important. Analysis might revolve around determining which factors (time or attributes) or events lead to the outcome by determining how the metrics differ between the groups. Conversely, if the cohorts were split based on an event type, questions may revolve around finding distinguishing outcomes (e.g., patients who took Drug A may have more strokes than patients who took Drug B). Exploration of cohorts that are split by time (e.g., the same patients over two different months) may be more open ended and require all metrics. The cohorts can be distinguished by time factors, event attributes, or events themselves (sequences of events or outcomes).

Results from purely statistical methods can also be difficult to parse and understand. Users may have different priorities and questions, which require different methods for sorting the results. For example, a user may be interested in *any* difference between the datasets, regardless of the direction of the difference, whereas another user may be interested only in results that occur more frequently in Cohort A. Integrated interaction techniques would allow users to specify their priorities when viewing results.

Purely visual tools for temporal event sequences are a good starting point for developing analysis tools for cohort studies but can be improved by the inclusion of the statistical tests used in automated approaches. For example, EventFlow assumes that each patient record consists of timestamped point events (e.g., heart attack, vaccination, first occurrence of symptom), temporal interval events (e.g., medication episode, dietary regime, exercise plan), and patient attributes (e.g., gender, age, weight, ethnic background).

In multiple case studies with EventFlow, the researchers repeatedly observed users visually comparing event patterns in one group of records with those in another group. In simple terms, the question was: what are the sequences of events that differentiate one group from the other? A common aspiration is to find clues that lead to new hypotheses about the series of events that lead to particular outcomes, but many other simple questions also involved comparisons. Epidemiologists analyzing the patterns of drug prescriptions [Monroe et al. 2013b] tried to compare the patterns of different classes of

drugs. Researchers analyzing task performance during trauma resuscitation [Carter et al. 2013] wanted to compare performance between cases where the response team was alerted of the upcoming arrival of the patient or not alerted. Transportation analysts looking at highway incident responses [Guerra-Gómez et al. 2011] wanted to compare how an agency handled its incidents differently from another. Their observations suggest that some broad insights can be gained by visually comparing pairs of EventFlow displays (e.g., users could see if the patterns were very similar overall between one month and the next) or very different (e.g., a lot more red or the most common patterns were different), but users repeatedly expressed the desire for more systematic ways to compare cohorts of records.

Our contribution is to enable researchers to be far more flexible in examining cohorts and facilitate human intervention where it can save time and effort. Because of the predefined problem space of comparing temporal event sequences, we can save users time by having answers to common questions readily available and giving them a starting point for their exploration. It is important to note that CoCo is intended for exploratory data analysis, which will reveal areas of interest to analysts and generate hypotheses. Users are encouraged to conduct follow-up (and more controlled) studies after they have identified possible hypotheses—such as clinical trials in the medical domain or A/B testing in the e-commerce domain.

Throughout this article, we use sample data of patients as they are transferred through a hospital. The patients are compared based on their outcome: those who lived versus those who died. All 2,356 patients are brought to the emergency room and contain the "Emergency" event and all patients are discharged with the "Exit" event. Other events include being admitted to a normal hospital room ("Normal Floor Bed"), being admitted to the ICU ("ICU"), and being administered aspirin ("Aspirin"). This dataset is a demonstration dataset only and uses no real data.

4. METRICS FOR COMPARING COHORTS

Metrics for comparing cohorts are numerous and can be grouped into five main categories: summary metrics, time metrics, event sequence (both whole record sequences and subsequences thereof) metrics, event attribute metrics, and record attribute metrics. These metrics are a direct result of observing EventFlow users as they analyzed cohorts of event sequences in nine case studies performed over 4 years [Monroe 2014]. Seven case studies were in the health care domain (with pharmacists and epidemiologists), one in sports analytics (basketball), and one in transportation.

4.1. Summary Metrics

Summary statistics apply to the cohorts as a whole and provide a high-level overview of the datasets.

Number of records. Total number of records in each cohort.
Number of events. Total number of events in each cohort.
Number of unique records. Total number of unique records in each cohort based on the sequence of events (absolute times are not considered).
Number of each event. Total number of occurrences for each event type per cohort.
Minimum, maximum, and average length of records. The length of a record is considered as the number of events in that record.

4.2. Event Sequence Metrics

Event sequence metrics deal with the order and structure of event sequences. Sequences are differentiated by whether they compose the entire record's history (*sequence*) or a *subsequence* thereof. Each of the following metrics can be presented as the

percent of records containing the event or sequence or as the percent of all events or sequences. The former method provides a sense of how many individual records had this sequence, whereas the latter method provides a sense of how events or sequences might repeat themselves within one record. For example, if an event occurs a high number of times but only in a few patients, it is implied that the event occurs frequently *per record*.

Prevalence of an event. The percent of records that contain an event.
Prevalence of a subsequence. The percent of records in which a subsequence appears. For example, the patients who lived are given aspirin before going to the emergency room more often than the patients who died.
Prevalence of a whole sequence. Percent of records with a given sequence.
Order of sequential events in a subsequence. The percent of records containing event A directly preceding event B versus B preceding A.
Commonly co-occurring (nonconsecutive) events. The percent of records containing both events A and B (in any order, with any number of events between them).
Prevalence of outcomes. A single event is prevalent as an "outcome" (i.e., the last event in the sequence). This metric in particular applies only to cohorts that are not already split on an outcome event.

4.3. Time Metrics

Time metrics deal with the timestamps at both the event and sequence levels—relative and absolute.

Absolute time of an event. Prevalence of a particular timestamp of an event or multiple events (e.g., all events in one cohort occurred on the same day).
Duration of interval events. The duration of a particular interval event. For example, this can be the length of exposure to a treatment or the duration of a prescription.
Duration between sequential events. The time between the end of one event and the beginning of the next, for example, the average length of time between hospital patients entering the emergency room and being transferred to the ICU.
Duration between co-occurring (nonsequential) events. The length of time between nonconsecutive events (two events with some number of other events occurring between them).
Duration of a subsequence. The length of time from the beginning of the first event in a subsequence to the end of the last event in the subsequence.
Duration from a fixed point in time. The length of time from a user-specified, fixed point—aligned by either a selected event or absolute date/time.
Duration of overlap in interval events. The overlap (or lack thereof) of interval events.
Cyclic events and sequences. The duration between cyclic events and sequences.
Survivor analysis. Rate at which an event or sequence occurs or diminishes over time.

Statistics for each of these metrics include the minimum, maximum, median, and average durations or values. Other summary statistics are also applicable.

4.4. Event Attribute Metrics

Any of the aforementioned metrics can be broken down by values of an attribute of the events instead of the event type itself. This can be done by swapping an event type by the values of a particular attribute. For example, in a medical dataset, we might be interested in seeing how a particular emergency room doctor might be related to the outcome of a patient. We would then switch all events of type "Emergency" with the value of its "doctor" attribute. If there are three doctors, this would create three new

pseudo-event types. We can use the metrics from earlier to analyze the difference in event sequences, times, or prevalence of each doctor in either cohort.

4.5. Record Attribute Metrics

Record-level attributes (such as patient gender or age) compare the cohorts as population statistics. General statistics across the entire dataset are a problem already tackled by analytics tools such as Spotfire [TIBCO 2014] or Tableau [Software 2014]; however, these tools look at a single attribute. For example, they might compare the number of males versus females or patients on Wednesday versus Thursday. There may be implications about the *combinations* of record attributes (e.g., the women on Wednesday vs. the women on Thursday vs. the men on Wednesday vs. the men on Thursday). In clinical trials, it is important that all patient attributes are balanced and currently no visual analytics tools exist for visually confirming that all attribute combinations are balanced.

4.6. Combining Metrics

The number of metrics is further multiplied because any combination of the aforementioned metrics is a new metric. For example, a sports analytics researcher may be interested in how a particular player (as an attribute of an event) performs within 2 minutes (time) after halftime (event order).

5. DESIGN GUIDELINES FOR HVHT VISUAL ANALYTICS TOOLS

Results from previous work [Malik et al. 2015] illustrated that expert analysts feel limited by the inability of current tools to evaluate the numerous hypotheses that can be applied to event sequences comparisons. Using a Design Methodology process [Sedlmair et al. 2012], we aimed to address these limitations and understand how to leverage the benefits of automated statistical analysis with user-guided exploration into a visual analytics tool.

We conducted interviews with three analysts experienced with event sequence visualization: a medical researcher from a local hospital, a graduate student at the University of Maryland, and a business school professor. All have used EventFlow [Monroe et al. 2013a] extensively and have active research projects comparing cohorts of patients. An initial version was implemented. After our three users had used the initial version with their own data and analytic goals, we interviewed them during a period of a month to collect feedback on the benefits and pitfalls of the initial version and analysts' needs when reviewing hypothesis results. Feedback was also collected from a dozen other short-term detailed demonstrations to potential future case study users, including a transportation analyst, business insight analysts, and statisticians.

Through this methodology, we designed and evaluated a visual analytics tool, CoCo, and distill our lessons learned into six design guidelines for balancing automated high-volume hypothesis testing with integrated visualization and interaction:

Guideline 1. Reduce wait times during computation.
Guideline 2. Convey hypotheses succinctly.
Guideline 3. Visualize statistical results and differences.
Guideline 4. Allow flexible methods for organizing results.
Guideline 5. Provide flexible interactions for parsing results.
Guideline 6. Provide context to highlight statistical uncertainty, false positives, and error rates.

Guideline 1: Reduce Wait Times During Computation

In any form of HVHT, wait times are a given, but this problem is especially prevalent when dealing with groups of event sequences because of the exponential number of unique sequences that exist in a single dataset. Consider the simple case of a dataset with only two events: A and B. Without considering repetitions, there are five unique event sequences that can occur:

$$A \qquad B \qquad A \rightarrow B \qquad B \rightarrow A \qquad AB \text{ (at the same time)}$$

When allowing repetition, the number of event sequences becomes infinite:

$$A \rightarrow A \qquad B \rightarrow B \qquad A \rightarrow A \rightarrow B \qquad AB \rightarrow B \qquad \ldots$$

Further, each event sequence can have multiple metrics applied to it. For example, with the sequence $A \rightarrow B$, we can consider the prevalence among records (i.e., percent of records containing this sequence), frequency (i.e., average number of occurrences per record), and duration (i.e., average time from A to B). When comparing cohorts, the application of each of these metrics to each cohort is equivalent to a hypothesis. Does $A \rightarrow B$ occur similarly in both cohorts, or does it occur significantly more in one than the other? Is the duration of $A \rightarrow B$ the same in both cohorts, or is it longer in one than the other? Thus, a simple dataset with only five event types can have hundreds of hypotheses applied to it.

In CoCo, we applied two different methods for computing the hypothesis tests: (1) performing all calculations ahead of time and providing results only when all results are complete and (2) calculating hypothesis tests by category (e.g., single event frequency, sequence frequency, time gaps, etc.) and allowing users to see results as they are available.

The first method results in long wait times but allows the results to be ranked in a more meaningful way. That is, by waiting for all results to be completed, the most "differentiating" or significant results can be displayed first, thus offering more guidance to the user about which results are important.

The second method allows users to see partial results as soon as they are ready. When a metric is fully calculated, the user can select that metric to see results for that metric. In early case studies, this enabled users to narrow their focus, although they found that they weren't necessarily interested in specific metrics, just the most major differences—regardless of metric type. We structured the metrics to first calculate the most simple metrics first (e.g., single event metrics), which enabled users to understand their datasets on a broader level before going into detailed sequence metrics.

Our future goal is to minimize wait time but give prompt feedback on which metrics might be meaningful to look at immediately, in accordance with Stolper et al.'s design guidelines for progressive visual analytics [Stolper et al. 2014].

Guideline 2: Convey Hypotheses Succinctly

In an initial implementation, we used the LifeLines2 [Wang et al. 2010] triangle scheme to display event sequences and organized results by metric (e.g., all results dealing with the occurrence of sequences were grouped together; all results dealing with co-occurrences of events were grouped together, etc.). With this organization scheme, the metric selected by the user implied a lot about the sequences in its result set, and all sequences looked identical. In feedback on this design, many users felt that only visualizing the sequence (with no indication of what the hypothesis was) was confusing and they would often have to remember which metric was selected.

We conducted interviews with three domain experts to determine how to distinguish between various event sequence features. In the interviews, we asked each expert how they would visually differentiate the following four types of sequences:

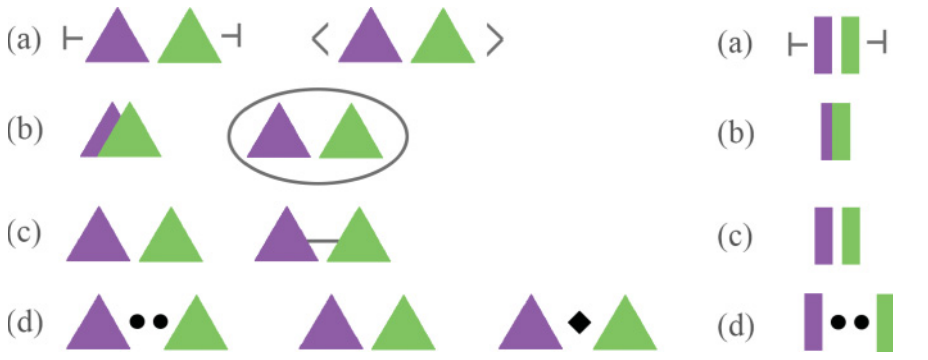


Fig. 2. Mockups of expert users' responses (left) and resulting glyphs (right) for visually differentiating four properties of event sequences: (a) whole record sequences, (b) concurrent events, (c) consecutive sequences, and (d) nonconsecutive sequences.

(a) Whole record sequence
(b) Concurrent events
(c) Consecutive sequences
(d) Nonconsecutive sequence

Mockups of responses are shown in Figure 2(a).

(a) Whole record sequences. A whole record sequence is a sequence that is an entire record (e.g., patient) history. Users suggested differentiating these sequences by adding markers indicating the beginning and end of the sequence, to signify no events occur before or after the sequence. Square markers were chosen over the angled brackets to avoid ambiguity with the notion of a "set."
(b) Concurrent events. Concurrent events are two or more events that occur at exactly the same timestamp. Users conveyed concurrent events by either overlapping them or grouping them with a circle. The overlapping method was preferred because it was more compact.
(c) Consecutive and (d) nonconsecutive sequences. A consecutive sequence is a series of events that appears without any other events within them. Nonconsecutive sequences may contain other events within them. Users had more variation in how they chose to differentiate consecutive versus nonconsecutive events. Two users chose to keep consecutive sequences the same while differentiating nonconsecutive sequences by placing a marker between events. The third user suggested the opposite: show no differentiation between nonconsecutive events but place a bar to join consecutively occurring events.

In a later version, we changed the event icons from triangles to slim rectangles to conserve screen real estate (and give users the option to toggle between the two versions). The current scheme is shown in Figure 2(b).

Guideline 3: Visualize Statistical Results and Differences

In designing the result displays, we needed a design that conveyed information about the difference in value (both magnitude and direction) and the statistical significance of the result. Additionally, we needed to avoid using color because color is already used to encode the event categories. We considered the three methods of visual comparison

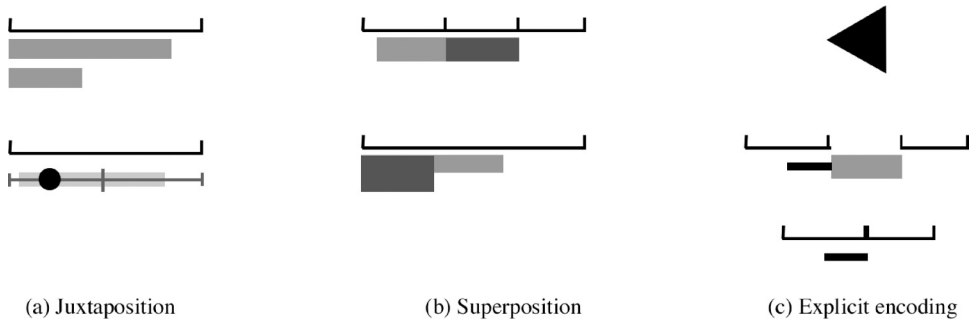


Fig. 3. Designs considered for presenting difference results between two cohorts: (a) juxtaposition (directly comparing two bars), (b) superposition (overlaying bars darkened area is the shared amount while the lightened area indicates the difference). The results are either side by side (top) or stacked (bottom), and (c) explicit encoding only, which encodes only information about the direction and magnitude of the difference, using shape and size.

outlined by Gleicher et al. [2011] to encode this data: (1) juxtaposition, (2) superposition, and (3) explicit encoding. Figure 3 shows the designs we considered.

(a) Juxtaposition showed the absolute values in each cohort and worked well for values that had a fixed range (e.g., percentages for 0% to 100%). However, it was not adaptable for variable-range values (e.g., time, where a difference can be as small as 1 minute or as large as 3 months) or for displaying time and prevalence metrics in the same view. It is also not ideal for scanning for differences easily because the difference is not explicitly encoded.
(b) Superposition has the advantage of displaying the raw values and direction of difference more clearly but had similar problems to juxtaposition in displaying time and prevalence results on the same axes; because it is axis dependent, it is not possible to display time and percentage in the same view.
(c) We found that an explicit encoding only offered the best option by allowing users to easily see and interpret differences between the datasets, despite that absolute values in each cohort are obscured. The absolute value information was available using interactions such as hover or details on demand to display them.

With the explicit encoding method, we were also able to explore different methods for encoding the differences: absolute difference, relative difference, and ratio. The values in datasets can be categorized in three groups. Take, for example, the occurrence of a sequence:

(1) Occurs in both datasets the same way (no difference)
(2) Occurs in both datasets, but more in one
(3) Occurs in only one dataset

Again, providing the absolute difference is sufficient when presenting prevalence results, because percentages are bounded to 100%. However, with results dealing with time, a single scale does not accurately convey differences because (1) time is unbounded, and (2) simply scaling the axis does not always work because even within a single dataset, different time granularity may exist (e.g., a hospital stay is on the order of days, whereas a prescription is on the order of months). Relative differences and ratios eliminate the problems of multiple units and granularities; however, we found that users understand ratios more clearly than relative differences. For example, it is easier to interpret "hospital stays are two times longer in cohort A than cohort B" rather than "hospitals stays are 100% longer." Because ratios can be anywhere from 1

(in case 1) to infinity (in case 3), we bound the axis to a user-defined maximum (default: 4×). If the ratio is above the maximum, the bar grows off the side, and if it is infinite, an infinity symbol is displayed next to the ratio bar.

Guideline 4: Allow Flexible Methods for Organizing Results

We explored four methods for organizing the results, each with its own benefits:

(1) Metric hierarchy. This was the approach taken in the initial version and it worked well in guiding the users. Users typically started their analysis based on sequence length, looking at single events before looking at longer sequences, then progressing based on their specific questions. This method worked best when users had specific questions about the datasets (e.g., if they were only concerned with whole record sequences).
(2) Flattened. We found that in open-ended and unstructured exploration, the users do not seem to care about what the metric is, just how important or distinguishing the result is. A flat design displays all hypothesis results in a single list view, regardless of metric or sequence type, and orders them by the significance and magnitude of difference.
(3) Sequence. Some researchers may have questions about a specific sequence of events. For these questions, it is best to group results by event sequence.
(4) Metric/flat list hybrid. In this view, the top 10 results for each metric are displayed. A hybrid view will give a good overview of the most important features of the dataset.

In our initial implementation, we organized results based on their category (Method #1). However, interviews with domain experts and users indicated that in their analyses, they didn't always care which metric was significant; they wanted all the significant results in one place, regardless of what type of metrics they corresponded to. We found that Method #2 is the most flexible for most uses and that users can use filters (Section 3) if they have specific questions dealing with a particular sequence or metric.

Guideline 5: Provide Flexible Interactions for Parsing Results

Displaying large result sets presents challenges in parsing them. We provide three interaction techniques for parsing the results. First, with so many hypotheses, not every hypothesis will apply to the dataset or domain. Researchers might have different priorities based on questions they already have. For example, some users might only be concerned with whole record sequences, while others want to see patterns across shorter subsequences. Some users might be concerned with only metrics dealing with prevalence, whereas others are interested in both time and prevalence metrics. Filtering and sorting provide flexibility by allowing users to manage their data based on what is relevant to their questions.

Second, as users sort through the results, they might easily disregard some hypothesis given their domain knowledge (e.g., results that are spurious correlations) and would need some way to keep track of everything they care about or have hidden. For this, we allow simple journaling options: starring, hiding, and annotating.

Lastly, when there are thousands of tested hypotheses, it is difficult for users to keep track of how many hypotheses they have viewed, how many are left to view, and of those results that are unviewed, which are significant. We solve this with a progress bar that indicates users' progress through the result set, so users feel comfortable that all possibly meaningful results have been reviewed.

Filtering is possible by:

—Event type (Figure 4(a)). Users can use the legend to select or deselect events they are interested in. Any sequences containing deselected events will not be shown.
—Record coverage (Figure 4(b)). Each sequence is displayed as a dot on a scattergram to display the number of records in each cohort that contains the sequence: the x-axis is cohort 1 and the y-axis is cohort 2. Users can filter based on the sample sizes to exclude results with very low or very high sample sizes. This can be used as a method for quality control (e.g., removing results with an insufficient sample size) or as a method for segmenting the results into more manageable pieces. For example, users may want to evaluate more frequent sequences first (e.g., sequences with 50% or more record coverage) before moving viewing less frequent sequences.
—Significance (Figure 4(c)). Users can filter based on three p-value groups: ≤ 0.01, ≤ 0.05, and > 0.05.
—Metric or sequence type (Figure 4(d)). Users can choose to see only time or prevalence metrics. Similarly, users might be interested in only single events, whole record histories, or partial subsequences.

Users can **sort** the results based on what they find most important (Figure 4(e)):

—Ratio and significance. Sort first by the significance level (p-value in three groups: ≤ 0.01, ≤ 0.05, and > 0.05.), then within each group, by the magnitude of the difference (descending). This is the default sorting option.
—Significance only. Sorted by the raw p-value (descending).
—Ratio only. Sorted by the absolute ratio (ascending or descending)
—Cohort 1 value. Sorted by the absolute value in alpha (descending).
—Cohort 2 value. Sort by the absolute value in beta (descending).

Guideline 6: Provide Context to Highlight Statistical Uncertainty, False Positives, and Error Rates

As users progress through the result set, it is difficult to understand if a result is meaningful based on a single result, especially when dealing with event sequences. For example, if patients visit the ICU after the emergency room more often in a cohort of patients who died versus lived, it may only be significant because the "ICU" event only occurs more often in the cohort of patients who died, so the added information of the "Emergency" event is not helpful. Additionally, providing overviews of the distribution of p-values allows the user to make more informed decisions about the significance of particular results and which results are valid. We provide context by providing details on demand, providing an overview of p-value distribution, and providing an overview of the users' progression through the result set.

Details on demand. Users are able to see the underlying data for a selected result. Depending on the type of metric, users will see different information. Because metrics dealing with prevalence are only a matter of percentage, all this data is shown in the result snapshot and the details on demand don't show any additional information. For metrics that show an average (e.g., all time metrics and frequency metrics), the details on demand show the exact distribution for all values (Figure 5). Additionally, the details on demand show high-level statistics about the distribution: sample size (n), average, minimum, maximum, and standard deviation.

P-value distribution. Based on the suggestions from two statisticians, we display the distribution of p-values across the various metrics. Using a simple table (Figure 6), users can understand whether the resulting distribution of p-values matches their expectations for meaningful results. Using a simple table, users can gain an understanding of the distribution of p-values and put a specific p-value into context (as

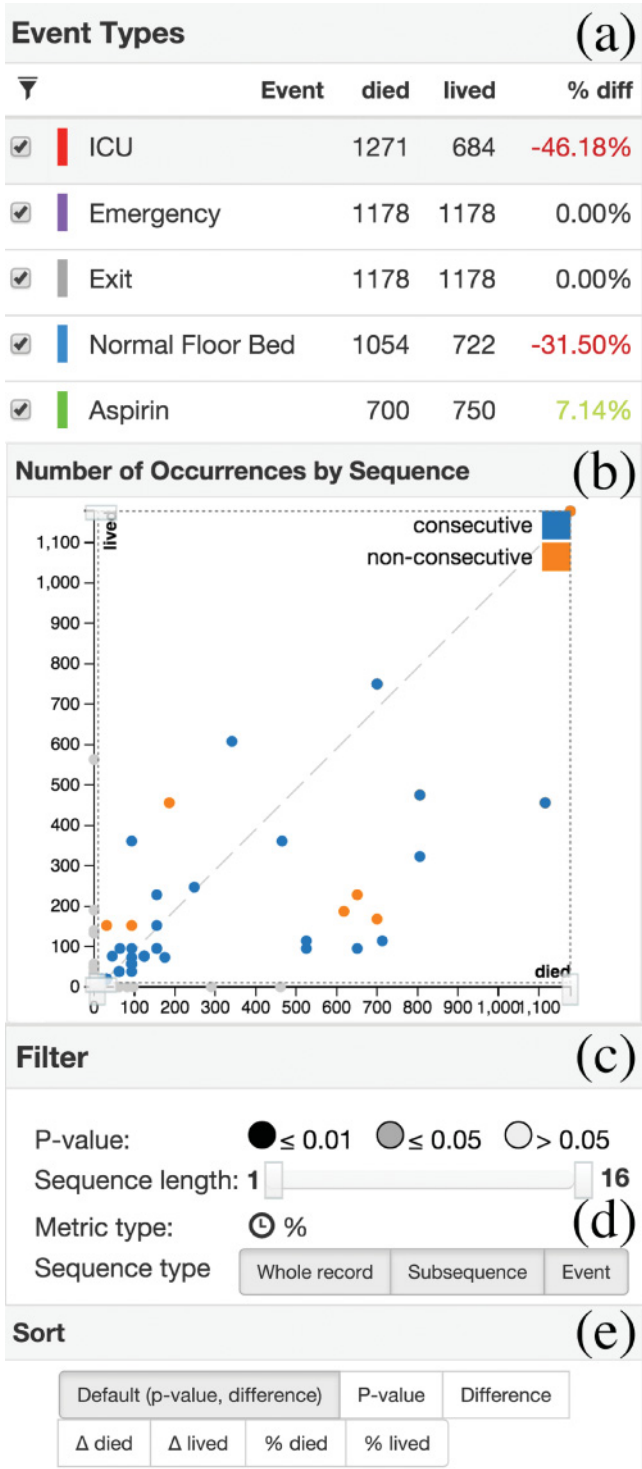


Fig. 4. Methods for sorting and filtering.

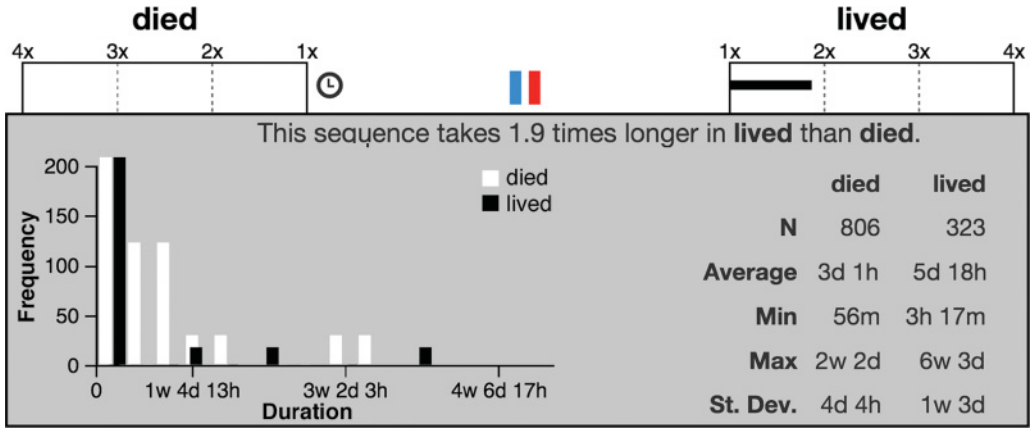


Fig. 5. Users can view details about a result by clicking it. Results that correspond to comparing averages (such as average duration or average frequency) will show the distributions of all the values and statistics about the average, minimum, maximum, and standard deviation in both cohorts.

	% Prevalence			Time	Frequency
≤ 0.01	14	2	27	4	2
≤ 0.05	4	1	6	0	0
> 0.05	6	4	81	3	1

Fig. 6. A simple table displays the number of results at each significance threshold by metric type.

p-values are heavily dependent on sample size). This information, combined with the other filtering controls (e.g., filtering sequences by a minimum required sample size), helps users reduce false positives and errors.

Progression. When reviewing a large list of results, it is unclear to users when everything has been reviewed, especially when they use filtering methods to view smaller pieces of the results at a time. A simple progress bar at the right of the results (Figure 7) shows the users' progress through the result set. It is a heatmap where each result is a single line and color indicates the following:

—Gray: result has been reviewed.
—Red: result has been calculated and has not been reviewed.

To make it more obvious that the user has not missed potentially significant results, we also encode the p-value in the heatmap using a colored border:

—Black indicates a p-value ≤ 0.01.
—Gray indicates a p-value ≤ 0.05.
—White indicates p-value > 0.05.

The progress bar serves as the scroll bar and minimap for the result set. Users can page through the data by scrolling along the progress bar. A thickened border indicates

Fig. 7. A progress bar indicates the users' progress through the result set. Gray indicates reviewed items and red are not viewed. The border corresponds to the significance of the result (black is $p < 0.01$, gray for $p < 0.05$, and white otherwise).

the portion of the data that is currently being viewed. The order in the progress bar matches the order of the detailed results and is determined by the user, based on the sort options provided.

6. CASE STUDIES

To investigate the strengths and limitations of CoCo as an automated cohort comparison tool, we conducted case studies following the procedure of a Multidimensional, Long-term In-depth case study (MILC) [Shneiderman and Plaisant 2006] with medical researchers from three institutions. Two were performed with a preliminary version of CoCo. The third was performed with the most recent version of CoCo, which was revised based on feedback from the first two case studies.

6.1. Exploring Adherence to Advanced Trauma Life Support Protocol

We worked with medical researchers at Children's National Medical Center who were investigating trauma teams' adherence to the Advanced Trauma Life Support (ATLS) protocol and possible reasons for deviations. One analyst was a clinician and researcher, and the other was a statistician. In a previous study [Carter et al. 2013], they found that about 50% of resuscitations did not follow the ATLS protocol. As a follow-up, the researchers' were interested in the following questions:

(1) What percent of patients are treated in adherence to protocol?
(2) Are there distinguishing attributes (e.g., time of day, patient gender, team lead) between protocol adherence and nonadherence?
(3) What are the most common deviations from the protocol?

After an initial training session and interviews to understand the researchers' goals and questions, we observed the researchers as they conducted a 3-hour session of data exploration and analysis.

The dataset consisted of 181 patient records, with event types for the five steps in the ATLS protocol: airway evaluation, listening for breath sounds, assessment of circulation, evaluation of neurological status disability, and temperature control. Patient attributes included injury severity score (ISS), the day of the week, length of hospital stay, time between notification and arrival at the hospital, and if the patient was admitted to the hospital.

The dataset was stored as a single file. The researchers used EventFlow's "group by attribute" feature to split the dataset into separate cohorts based on attributes and adherence to the protocol. Over the course of the 3 hours, the researchers split the dataset six ways to load six different pairs of cohorts in CoCo as they explored different hypotheses:

(1) Patients treated in adherence to the ATLS protocol versus those that showed any deviation
(2) Patients admitted to the floor versus ICU (with discharged patients removed)

(3) Patients who arrived with at least 5 minutes' warning before arrival at the trauma bay versus those who arrived with no warning ("now" patients)
(4) Patients with a high (above 25) versus low ISS
(5) Patients treated on the weekend versus on a weekday
(6) Patients treated during the day versus at night

In every comparison, the analysts began by narrowing the results by metric. The analysts started by looking at the prevalence of single events to determine how often they occurred. The analysts then looked at the most differentiating *entire record* sequences, because the subsequences were less informative about how the protocol was followed. They would then make their way down the provided metrics list, in the order that they appeared: most differentiating time gaps and then prevalence of record attributes. They did not look at the prevalence of record attribute combinations for any of the datasets.

For this dataset, the analysts expected to see that all records contained every event. This finding was not observed for two of the comparisons: correctly treated patients versus those with deviations and day versus night patients, with the latter of both groups receiving the airway check significantly less than daytime patients. In the day versus night group, the analyst also found that the "most differentiating sequence" was the *correct* order, meaning that the nighttime patients were treated in the correct order significantly less often than daytime patients. Additionally, patients treated at night had more variance in the procedure, with 26 unique sequences in the 83 nighttime patients versus 20 unique sequences in the 101 daytime patients. A possible reason for this finding is that during the day, nurse practitioners perform these procedures, but at night, less experienced junior residents are on call instead.

At times, the researchers saw that certain groups occurred only rarely in the cohorts (fewer than 20 times), so the researchers decided not to consider the comparisons. For example, among patients admitted to the ICU or floor, only about 80 patients remained, making the sample sizes too small to run many of the significance metrics about event types. As one analyst worked to confirm her expectations and check several hypotheses, she found a surprising and potentially important result: about 25% more patients who were admitted to the floor were "now" patients ($p < 0.05$), which led to splitting the cohort into the third group: "now" versus "not now" patients.

In the closing interview, one analyst said, "We don't need to solve everything with EventFlow and CoCo. These tools let us explore the data and narrow our hypothesis." From these results, the analysts submitted abstracts about and presented these findings at an internal symposium on trauma care. This first case study suggests that CoCo can be effective for exploratory analysis and hypothesis generation.

6.2. Comparing Algorithms for Distinguishing Types of Radiation to the Bone

We also worked with pharmacists at the Department of Pharmaceutical Health Services Research at the University of Maryland School of Pharmacy in Baltimore. In previous work, the researchers were interested in developing an algorithm using claims data to differentiate between radiation delivered to the bone versus radiation delivered to the prostate gland, because billing codes available in claims data do not distinguish the site of radiation. Reliable measures for identifying the receipt of radiation to the bone are important in order to avoid bias in estimating the prevalence and/or mortality impact of skeletal-related events, including radiation to the bone.

Studies using healthcare claims employ various claims-based algorithms to identify radiation to the bone and mostly condition on prior claims with a bone metastasis diagnosis (billing) code [Sathiakumar et al. 2013; Nørgaard et al. 2014; Lage et al. 2008]. They developed three classification algorithms that were compared in pairs

using CoCo and EventFlow to investigate the timing of possible radiation to the bone among patients diagnosed with incident metastatic and nonmetastatic prostate cancer. One algorithm was based on prior literature, while the other two were based on insights gained from data visualization software. Based on clinical input regarding the duration of palliative [Hartsell et al. 2005; Lutz et al. 2014] versus curative radiation, the researchers investigated the length of radiation episodes and found differences between cohorts in terms of the length of radiation. As expected, patients diagnosed with metastatic disease received a shorter course radiation than patients diagnosed with nonmetastatic disease.

The feedback on CoCo was positive and the team valued the opportunity to visually compare cohorts of patients using summary statistics that pertained to the timing and frequency of events. The graphical results were shared with clinicians on the research team in order to determine whether the patterns were consistent with their expectations. The researchers felt the meaning of metrics could be explained more clearly; it was sometimes unclear what the x-axis represented and what statistical tests were used. They also suggested always showing the event labels, particularly for single-event metrics, to make understanding the icons a bit easier. The researchers expressed a need to be able to sort the results with different factors, including by raw percentage of values in each cohort. We implemented this feature before the formal case study.

6.3. Medication Adherence Patterns of Hypertension Patients

Researchers at the University of Maryland's School of Pharmacy and School of Business are analyzing the medication adherence patterns of patients on diuretics (i.e., are patients taking their drugs as prescribed, in which combinations, what characterizes the gaps between prescriptions, etc.) [Bjarnadottir et al. 2015]. In particular, they are interested in the differences between high-cost versus low-cost diuretics patients and want to know what patterns are representative of each group. Researchers became overwhelmed as they sought to evaluate hypotheses that dealt with hundreds of unique event sequences. This case study provides an illustrative example of the challenges that researchers and analysts encounter and describes how the implementation of new visualization interaction techniques for event sequence hypotheses in CoCo enables the automatic analysis of two groups of records.

The data these researchers gathered consisted of prescription refill histories of five drug classes commonly used to treat hypertension. The data spanned 1 year and contained over 1 million patients. The data also included the total cost of all prescriptions over the year. The researchers wanted to compare whether hypertension drug adherence affected the cost that patients incurred over a year. In other words: could taking medication as prescribed result in lower overall prescription costs?

Current methods for adherence analysis consist merely of calculating a Medication Possession Ratio (MPR) [Andrade et al. 2006] or similar aggregated measures that do not represent the diversity of patterns found in the data. The MPR for a predefined period is calculated as

$$MPR = \frac{\text{number of days prescribed}}{\text{days elapsed over period}}.$$

For example, if patients only filled one 30-day prescription over a period of 90 days, their MPR is 30/90, or 1/3. This method oversimplifies a patient's prescription history into a single number, which may not provide an accurate representation of his or her adherence behavior. A patient might refill prescriptions early when planning to leave on vacation, thus leaving a larger-than-usual gap in his or her prescription history, when in fact he or she was taking the medication regularly. Conversely, a patient who switches to another medication after a recent prescription refill may have a history that incorrectly indicates that the patient regularly took his or her prescription.

Table I. Number of Hypotheses Generated by Metric and Sequence Type

Metric	Sequence Type	Hypotheses
Record Coverage	Event	2
	Whole record	18
	Subsequence	64
	Co-occurring events	4
	Record attributes	2
Duration	Event pairs	2
Frequency	Event	2
Total		**94**

6.3.1. Data Processing. We report here only on the analysis of the adherence patterns of patients who took medications from only one drug class: diuretics, which consisted of a total of 113,401 patients. The dataset consisted of two event types: diuretic and gap, where diuretic indicated the start time of a prescription and gap indicated the start time of no medication usage. The patients were categorized into HIGH- versus LOW-cost patients based on the distribution of prescription costs for the patients. Patient costs ranged from $0 to $9,528 (USD). Most patients (55%) had no prescription costs and the average cost was $25.39. We excluded patients with $0 costs and patients with more than $380 (top %1), to exclude outliers.

The final datasets consisted of 3,958 patients (21,066 events, 720KB) categorized as HIGH cost and 38,175 patients (144,433 events, 5MB) categorized as LOW cost.

6.3.2. Comparison Metrics. The metrics used were based on a subset of the taxonomy presented in Section 4, where the following terms are defined:

> *Whole record sequence*. A set of two or more events following one another that compose an entire record history.
> *Subsequence*. A set of two or more events following one another that compose a piece of a record history.
> *Consecutive*. A sequence of events that occur in uninterrupted succession.
> *Concurrent*. Two or more events that occur at the same moment in time.

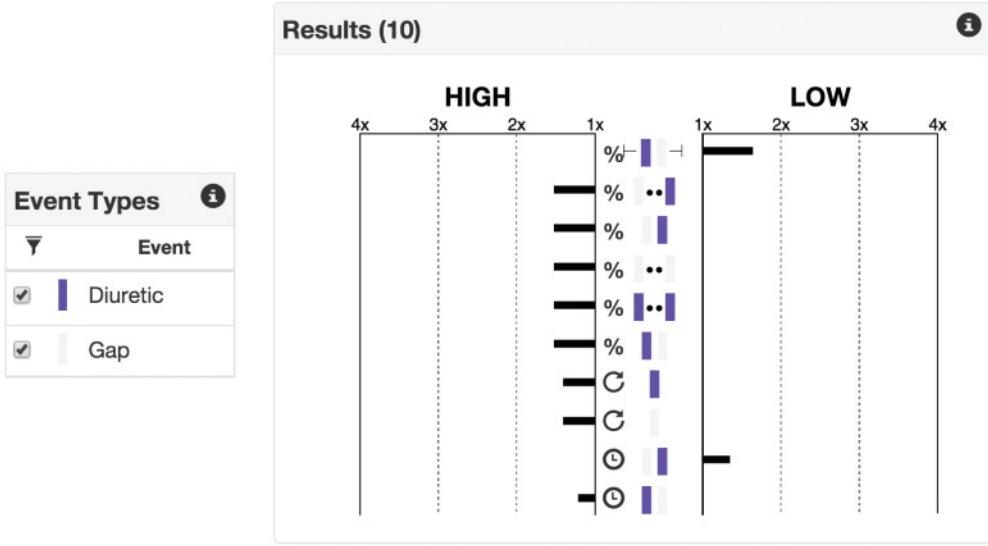
Specifically, the researchers were interested in:

—Record coverage of events, subsequences, and whole records
—Duration of co-occurring events
—Frequency of single events
—Occurrences of record-level attributes

6.3.3. System Use and Results. The final version of CoCo was used to compare prescription patterns of high- versus low-cost patients. In total, CoCo generated results for a total of 94 hypotheses. The hypotheses are broken down by metric and sequence type in Table I. The analyst was a professor at the University of Maryland's School of Business and was assisted in using CoCo by a graduate researcher.

The analyst first used the **Sequence Occurrence** panel to review only results with a sufficient sample size. The threshold was set at 10% of each cohort, or 395 in the HIGH-cost group and 3,817 in the LOW-cost group, which reduced the number of hypotheses to review to 24. Next, the remaining insignificant results ($p > 0.05$) were removed using the **Filter by Significance** feature, leaving a more simplified display of 21 results.

Finally, the results were **Filtered by Sequence Length**, to view only sequences of length 1 (single events) or 2 (event pairs). Because there are only two event types in the



(a) The final hypotheses results from the medication adherence study. Patients are categorized as HIGH versus LOW cost. After filtering by significance, sample size, and sequence length, there were 10 results remaining. High-cost patients tended to have more gaps in their data, as well as longer and more frequent prescriptions. Low-cost patients had more instances of a single prescription and ceased using medication.

(b, above) The results were filtered by p-value (results with $p > 0.05$ were removed) and by sequence length (showing only single events and pairs of events).

(c, right) Results were filtered to only include sequences that occurred in 10% or more of records in each cohort.

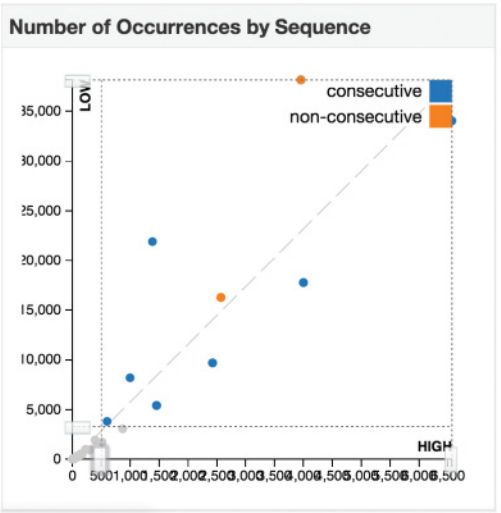


Fig. 8. Final results and usage of drug pattern case study. Analysts used the Sequence Occurrence panel (c) to control sample size, and the Filter panel (b) to control significance and sequence length. This resulted in only 10 hypotheses (a) for the researchers to manually review.

dataset, longer sequences were just repetitions of length 2 or less, so this was all that was necessary to view all unique patterns. Thus, there were 10 remaining hypotheses to review in detail. The final result display and settings are shown in Figure 8.

The analysts used the **Default Sorting**, then evaluated the remaining hypotheses one by one, using context information provided in the **Details on Demand Panel**. This made it easy to conclude that high-cost patients tended to have longer sequences, with more gaps and prescription refills, whereas low-cost patients had shorter

sequences, most commonly filling only a single prescription. Low-cost patients also had significantly longer gaps between prescription refills. As a follow-up, analysts will incorporate medical claims data to understand the more serious medical implications of low medication adherence (such as heart attacks or stroke).

6.4. Case Studies in Nonmedical Domains

Additional case studies and targeted controlled studies will be necessary to characterize the effectiveness of CoCo. We are working with researchers in other fields, including education (analyzing student enrollment patterns), web log analysis, and transportation data.

7. CONCLUSIONS AND FUTURE WORK

CoCo is a novel visual analytics tool with balanced integration of visual analytics and statistics. CoCo's benefits include better collaboration among colleagues, easier intermediate results discussion, and meaningful outcome presentations. The implementation and use of CoCo with domain experts uncovers six design guidelines that can be extended to other HVHT visual analytics tools. While we focus on case studies in the medical domain, these techniques can be used in any event sequence dataset, such as clickstreams or transportation reports with the inclusion of more metrics. More work would be needed to understand how more dense event streams (e.g., with clickstreams) would affect the comprehensibility of the results. We have begun case studies in the business and transportation domains.

CoCo's interface and visualizations are extendable in many ways. First, the metrics implemented are already proving valuable, but many more metrics are possible. Additional data mining and statistical techniques could be added to improve insight discovery, such as anomaly detection to find unusual records or clustering to find similar records across the datasets. More work would be required to understand how these controls could be more transparent to the user. It seems the interaction and visualization techniques proposed in this work could be applied to traditional data mining results (e.g., frequent patterns), but more research is required. As we continue developing CoCo, we will conduct controlled experiments to understand its strength and weaknesses, as well as long-term case studies with domain experts to demonstrate value with realistic problems and to guide our development.

We recognize that there are limitations to CoCo in terms of the complexity of datasets, current emphasis on two cohorts, and the need for more user control on which events to study. On the other hand, the fresh possibilities for statistical comparisons, supported by visual presentations and an intelligent user interface, open many doors for further research. We are encouraged by our initial feedback and we see many avenues for future research that will empower medical and other researchers as they conduct exploratory data analysis on temporal event sequences.

ACKNOWLEDGMENTS

The authors wish to thank our partners Ebere Onukwugha at the School of Pharmacy at the University of Maryland, Baltimore, and Randall Burd and Rachel Webman at Children's National Medical Center in Washington, DC. This work is supported in part by Adobe. We gratefully acknowledge funding provided by the University of Maryland/Mpowering the State through the Center for Health-related Informatics and Bioimaging.

REFERENCES

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM International Conference on Management of Data (SIGMOD'93)*. ACM, New York, NY, 207–216.

Susan E. Andrade, Kristijan H. Kahler, Feride Frech, and K. Arnold Chan. 2006. Methods for evaluation of medication adherence and persistence using automated databases. *Pharmacoepidemiology and Drug Safety* 15, 8 (2006), 565–574.

Stephen D. Bay and Michael J. Pazzani. 2001. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery* 5, 3 (2001), 213–246.

Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29, 4 (August 2001), 1165–1188. DOI: http://dx.doi.org/10.1214/aos/1013699998

Viv Bewick, Liz Cheek, and Jonathan Ball. 2004. Statistics review 12: Survival analysis. *Critical Care* 8, 5 (2004), 389–394. DOI: http://dx.doi.org/10.1186/cc2955

Margret Bjarnadottir, Sana Malik, Eberechukwu Onukwugha, Tanisha Gooden, and Catherine Plaisant. 2015. Understanding adherence and prescription patterns using large scale claims data. *Pharmacoeconomics* 34, 2 (Feb. 2016), 169–79. DOI: 10.1007/s40273-015-0333-4

Elizabeth Carter, Randall Burd, Megan Monroe, Catherine Plaisant, and Ben Shneiderman. 2013. Using eventflow to analyze task performance during trauma resuscitation. In *Proceedings of the Workshop on Interactive Systems in Healthcare* (2013).

David Collett. 2003. *Modelling Survival Data in Medical Research*, 2nd ed. Chapman and Hall/CRC Press.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association* 56, 293 (1961), 52–64.

Mathieu Dupont, Arnaud Gacouin, Hervé Lena, Sylvain Lavoué, Graziella Brinchault, Philippe Delaval, and Rémi Thomas. 2004. Survival of patients with bronchiectasis after the first ICU stay for respiratory failure. *Chest* 125, 5 (May 2004), 1815–1820.

Paolo Federico, Jürgen Unger, Albert Amor-Amors, Lucia Sacchi, Denis Klimov, and Silvia Miksch. 2015. Gnaeus: Utilizing clinical guidelines for knowledge-assisted visualisation of EHR cohorts. In *Proceedings of the EuroVis Workshop on Visual Analytics (EuroVA'15)*, Enrico Bertini and Jonathan C. Roberts (Eds.). The Eurographics Association. DOI: http://dx.doi.org/10.2312/eurova.20151108

Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D. Hansen, and Jonathan C. Roberts. 2011. Visual comparison for information visualization. *Information Visualization* 10, 4 (September 2011), 289–309. DOI: http://dx.doi.org/10.1177/1473871611416549

Manish K. Goel, Pardeep Khanna, and Jugal Kishore. 2010. Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research* 1, 4 (October 2010), 274–278. DOI: http://dx.doi.org/10.4103/0974-7788.76794

John Alexis Guerra-Gómez, Krist Wongsuphasawat, Taowei David Wang, Michael L. Pack, and Catherine Plaisant. 2011. Analyzing incident management event sequences with interactive visualization. In *Proceedings of the Transportation Research Board 90th Annual Meeting*.

Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. 2007. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery* 15, 1 (2007), 55–86. DOI: http://dx.doi.org/10.1007/s10618-006-0059-1

William F. Hartsell, Charles B. Scott, Deborah Watkins Bruner, Charles W. Scarantino, Robert A. Ivker, Mack Roach, John H. Suh, William F. Demas, Benjamin Movsas, Ivy A. Petersen, Andre A. Konski, Charles S. Cleeland, Nora A. Janjan, and Michelle DeSilvio. 2005. Randomized trial of short- versus long-course radiotherapy for palliation of painful bone metastases. *Journal of the National Cancer Institute* 97, 11 (2005), 798–804.

Maureen J. Lage, Beth L. Barber, David J. Harrison, and Sun Jun. 2008. The cost of treating skeletal-related events in patients with prostate cancer. *American Journal of Managed Care* 14, 5 (May 2008), 317–322.

Tim Lammarsch, Wolfgang Aigner, Alessio Bertone, Silvia Miksch, and Alexander Rind. 2014. Special section on visual analytics: Mind the time: Unleashing temporal aspects in pattern discovery. *Computer Graphics* 38 (February 2014), 38–50. DOI: http://dx.doi.org/10.1016/j.cag.2013.10.007

Guimei Liu, Mengling Feng, Yue Wang, Limsoon Wong, See-Kiong Ng, Tzia Liang Mah, and Edmund Jon Deoon Lee. 2011. Towards exploratory hypothesis testing and analysis. In *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering (ICDE'11)*. IEEE Computer Society, Washington, DC, 745–756.

Stephen T. Lutz, Joshua Jones, and Edward Chow. 2014. Role of radiation therapy in palliative care of the patient with cancer. *Journal of Clinical Oncology* 32, 26 (Sept. 2014), 2913–2919. DOI: http://dx.doi.org/10.1200/JCO.2014.55.1143

Nizar R. Mabroukeh and Christie I. Ezeife. 2010. A taxonomy of sequential pattern mining algorithms. *Computer Surveys* 43, 1, Article 3 (Nov. 2010), 3:1–3:41 pages. DOI: http://dx.doi.org/10.1145/1824795.1824798

Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. 2015. Cohort comparison of event sequences with balanced integration of visual analytics and statistics.

In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI'15)*. ACM, New York, NY, 38–49. DOI: http://dx.doi.org/10.1145/2678025.2701407

Megan Monroe. 2014. *Interactive Event Sequence Query and Transformation*. Ph.D. Dissertation. University of Maryland, College Park, MD.

Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. 2013a. Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (December 2013), 2227–2236.

Megan Monroe, Tamra E. Meyer, Catherine Plaisant, Rongjian Lan, Krist Wongsuphasawat, Trinka S. Coster, Sigfried Gold, Jeff Millstein, and Ben Shneiderman. 2013b. Visualizing patterns of drug prescriptions with EventFlow: A pilot study of asthma medications in the military health system. In *Proceedings of Workshop on Visual Analytics in Healthcare.*

Mette Nørgaard, Annette Østergaard Jensen, Jacob Bonde Jacobsen, Kara Cetin, Jon P. Fryzek, and Henrik Toft Sørensen. 2014. Skeletal related events, bone metastasis and survival of prostate cancer: A population based cohort study in denmark (1999 to 2007). *Journal of Urology* 184, 1 (October 6, 2015), 162–167. DOI: http://dx.doi.org/10.1016/j.juro.2010.03.034

Oracle. 2011. *Oracle Health Sciences Cohort Explorer User's guide*. Technical Report. Oracle. Retrieved from http://docs.oracle.com/cd/E24441_01/doc.10/e24437.pdf.

Adam Perer and Fei Wang. 2014. Frequence: Interactive mining and visualization of temporal frequent event sequences. In *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI'14)*. ACM, New York, 153–162. DOI: http://dx.doi.org/10.1145/2557500.2557508

Nalini Sathiakumar, Elizabeth Delzell, Michael Morrisey, Carla Falkson, Mellissa Yong, Victoria Chia, Justin Blackburn, Tarun Arora, and Meredith Kilgore. 2013. Mortality following bone metastasis and skeletal-related events among patients 65 years and above with lung cancer: A population-based analysis of U.S. Medicare beneficiaries, 1999-2006. *Lung India* 30, 1 (2013), 20–26. DOI: http://dx.doi.org/10.4103/0970-2113.106127

Michael Sedlmair, Miriah Meyer, and Tamara Munzner. 2012. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (December 2012), 2431–2440. DOI: http://dx.doi.org/10.1109/TVCG.2012.213

Juliet P. Shaffer. 1995. Multiple hypothesis testing. *Annual Review of Psychology* 46, 1 (1995), 561–584. DOI: http://dx.doi.org/10.1146/annurev.ps.46.020195.003021

Ben Shneiderman and Catherine Plaisant. 2006. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV'06)*. ACM, New York, NY, 1–7. DOI: http://dx.doi.org/10.1145/1168149.1168158

Tableau Software. 2014. Tableau. Retrieved from http://www.tableausoftware.com/.

Charles D. Stolper, Adam Perer, and David Gotz. 2014. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1653–1662. DOI: http://dx.doi.org/10.1109/TVCG.2014.2346574

TIBCO. 2014. Spotfire. Retrieved from http://spotfire.tibco.com/.

Katerina Vrotsou and Aida Nordman. 2014. Interactive visual sequence mining based on pattern-growth. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST'14)*. 285–286. DOI: http://dx.doi.org/10.1109/VAST.2014.7042532

Katerina Vrotsou, Anders Ynnerman, and Matthew Cooper. 2014. Are we what we do? Exploring group behaviour through user-defined event-sequence similarity. *Information Visualization* 13, 3 (2014), 232–247. DOI: http://dx.doi.org/10.1177/1473871613477852

Taowei David Wang, Krist Wongsuphasawat, Catherine Plaisant, and Ben Shneiderman. 2010. Visual information seeking in multiple electronic health records: Design recommendations and a process model. In *Proceedings of the 1st ACM International Health Informatics Symposium (IHI'10)*. ACM, New York, NY, 46–55. DOI: http://dx.doi.org/10.1145/1882992.1883001

Zhiyuan Zhang, David Gotz, and Adam Perer. 2014. Iterative cohort analysis and exploration. *Information Visualization* 14, 4 (Oct. 2015), 289–307.

Jian Zhao, Zhicheng Liu, Mira Dontcheva, Aaron Hertzmann, and Alan Wilson. 2015. MatrixWave: Visual comparison of event sequence data. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. ACM, New York, NY, 259–268. DOI: http://dx.doi.org/10.1145/2702123.2702419

Received July 2015; revised December 2015; accepted December 2015