Visual Cohort Queries for High-Dimensional Data: A Design Study

Wanchen Zhao, MSIS[1], David Borland, PhD[2], Arlene E. Chung[3,4], MD, MHA MMCi, David Gotz, PhD[1,3]
[1]School of Information and Library Science, [2]RENCI,
[3]Carolina Health Informatics Program, [4]School of Medicine
University of North Carolina, Chapel Hill, NC, USA

Abstract

The large collections of electronic health data gathered by modern health institutions are increasingly being leveraged as a source of real-world evidence within population health studies. Cohort selection is a critical first step in these studies. However, querying for patient data within complex medical databases can be challenging due to two key concepts: (1) the high-dimensionality of medical data, and (2) the temporal nature of many queries (e.g., "patients with a specific medical procedure within X days after diagnosis"). Visual interfaces which enable non-technical experts to define queries of this type are available in systems such as the widely used i2b2 platform. However, using such tools to retrieve a satisfactory cohort for a given study remains difficult, typically requiring users to employ an iterative cohort refinement process using multiple queries. This paper reports results from a formative design study aimed at gaining a better understanding of the iterative query process, identifying challenges faced by users as they define cohorts, and gathering feedback on a preliminary design for a novel interactive visual query interface.

1 Introduction

Cohort analysis is a foundational technique used within a vast range of health-focused research activities[1]. In this analytical method, data from groups of people are analyzed with respect to statistical outcome variables to help answer research questions. A well-designed study requires researchers to identify both (1) appropriate outcome variables (e.g., metrics which are appropriate to the research question), and (2) a high-quality cohort containing data about a representative set of patients to support the research activity.

In high-quality prospective studies such as randomized controlled trails (RCTs), inclusion and exclusion criteria for cohorts are carefully designed and the participants are typically randomly assigned to different groups of a study. This cohort design process is a critical part of a study, and the design choices have great impact on the validity and generalizability of the study results.

As health institutions amass ever larger and more complex collections of data during normal operations, similar cohort analysis approaches are increasingly being applied to conduct retrospective studies. This approach is supported by large investments in technologies designed to make the complex health data captured in these systems more accessible for research purposes. Such technologies include, for example, the NIH-funded i2b2 (Informatics for Integrating Biology and the Bedside) platform[2], as well as the OHDSI collaborative's suite of software tools and standards[3].

The key promise of retrospective analysis using real-world data from health systems is that the rich variation of patient experience—captured in the complex high-dimensional structure of electronic health data systems and densely sampled over time—will help provide more nuanced insights about real-world health than typical RCTs, which study more homogeneous populations. However, the very complexity of the data that makes such retrospective studies promising also makes it difficult for researchers to define high-quality cohorts. To retrieve data for a cohort, researchers express their inclusion and exclusion criteria as complex queries. The challenges in defining such queries come from several factors, including the high-dimensionality of the data (e.g., a single medical condition may be represented via a wide variety of diagnosis codes or lab test results), and the temporal nature of many study constraints (e.g., a treatment occurring within a certain time period after diagnosis).

A variety of visual query interfaces haven been developed to support this cohort selection process, including the rich query capability provided within the i2b2 system. Previous work on visual cohort query systems has shown that queries are often performed iteratively in a "trial-and-error" manner until a satisfactory cohort has been identified[4,5]. However, these tools often provide limited support for guiding users toward higher quality results.

In this paper, we describe a design study evaluating a new visual query interface that enables users to construct complex cohort definitions and supports iterative refinement of initial temporal queries. The design introduces (1) a flowchart-like representation of complex temporal query logic, (2) a novel "query time frame" concept to segment the

events of the returned cohort and facilitate query refinement based on events both ***within*** and ***outside of*** the initial query time frame, and (3) highlighting of relevant events in the context of the query based on an information gain metric. Based on interviews with three medical researchers who currently use the i2b2 query interface, we report findings regarding users' iterative query process, challenges faced during cohort construction, and feedback on our initial design.

2 Related Work

With the development of modern visual analytics techniques, a number of advanced visual query and analysis platforms have been developed. Generally speaking, these techniques have focused on two distinct phases of query construction. In the first phase, users define an initial query using a visual query interface. In the second phase, visualization-based representations help users explore the data returned by a query. Users can iterate between these two phases to perform query refinement.

2.1 Visual Query Interface

Visual query systems aim to provide intuitive yet expressive interfaces through which users can interactively define the inclusion and exclusion criteria that define a cohort of interest. Previous work has classified query representation into four major categories: form-based, diagram-based, icon-based, and hybrid (a combination of the other three types)[6]. Since the 1990s, form-based query representations[6] have very been widely used because they are relatively easy to implement and are well-structured[7]. However, form-based interfaces can also be less flexible than alternative designs. The i2b2 query system is an example of a form-based interface designed specifically for cohort queries: users define constraints using a form-based interface, and can define temporal relationships via sequential constraints between sections of the form. Diagram-based query representations are more flexible[4,8], and can effectively express complex query logic. However, larger diagrams can become difficult to interpret. In contrast, icon-based representations are often simpler to read but also less expressive. Given these tradeoffs, hybrid approaches are also common[9,10].

2.2 Query Result Visualization

Reporting tools of various kinds have long been used to depict the results returned by cohort queries. For example, various portions of i2b2 contain commonly used visualizations such as bar charts, pie charts, and various forms of line graphs.[9] These traditional chart types are very useful, but have trouble communicating variation over time within a cohort. For this reason, more specialized timeline or event sequence approaches have been adopted within multiple systems[11–15]. These approaches are effective at showing the sequential order of events in the cohort, with various solutions to help the visualizations scale to larger datasets[8,16–19]. Other approaches include Sankey diagrams[20] and network diagrams[21].

These visualizations present the results of a query for users to review, often with interactive exploratory capabilities that enable users to examine patterns and subsets of data within the returned cohort. When combined within a single platform with a visual query interface, these query result visualizations can indirectly support iterative query refinement providing the means for a user to better understand the qualities of a cohort returned by an issued query. One goal in the proposed design outlined in this paper is to more tightly integrate these elements to provide stronger support for iterative query refinement.

3 Design Study

We conducted a design study to evaluate the preliminary design of a new iterative cohort query interface. In the study, we designed a preliminary visual query interface, developed a functional prototype, and evaluated the prototype interface with real users as applied to sample query tasks inspired by users' day-to-day cohort query activities. The results highlight key benefits of the proposed approach and provide insights that can inform future visual query interface designs.

3.1 User Interface Design

Motivated by the iterative nature of the cohort query process, we developed a visual query interface design that directly supports query refinement. Toward this goal, we began the design process with three key requirements (R1-R3):

> **R1.** The visual query interface must provide intuitive easy-to-use representations of non-temporal query constraints and provide visualizations of the data that match a given query specification. We note that this is the typical design goal of most existing visual cohort query interfaces.

R2. The visual query interface must provide intuitive representations of temporal constraints, including complex nested dependencies between query criteria.

R3. The system should aim to directly inform users about promising ways to modify query specification to improve cohort quality. Moreover, the system should provide users with information about the impact that changes to the cohort constraints would have on the results returned by the query.

Preliminary Design. Motivated by the design requirements enumerated above, we developed a visual query interface design as illustrated in Figure 1 and Figure 2. The design incorporates four coordinated panels (an information panel, a query flow panel, a scatter chart panel, and a distribution panel) to visualize: the query constraints that define a query, various views of the data present within a queried cohort, and perhaps most critically, information about events occurring before, during, and after the time period specified in a query.

Information Panel. The information panel (Figure 1, panel 1) contains general information about the cohort results such as the number of patients returned by a query and the number of events occurring for those patients within the

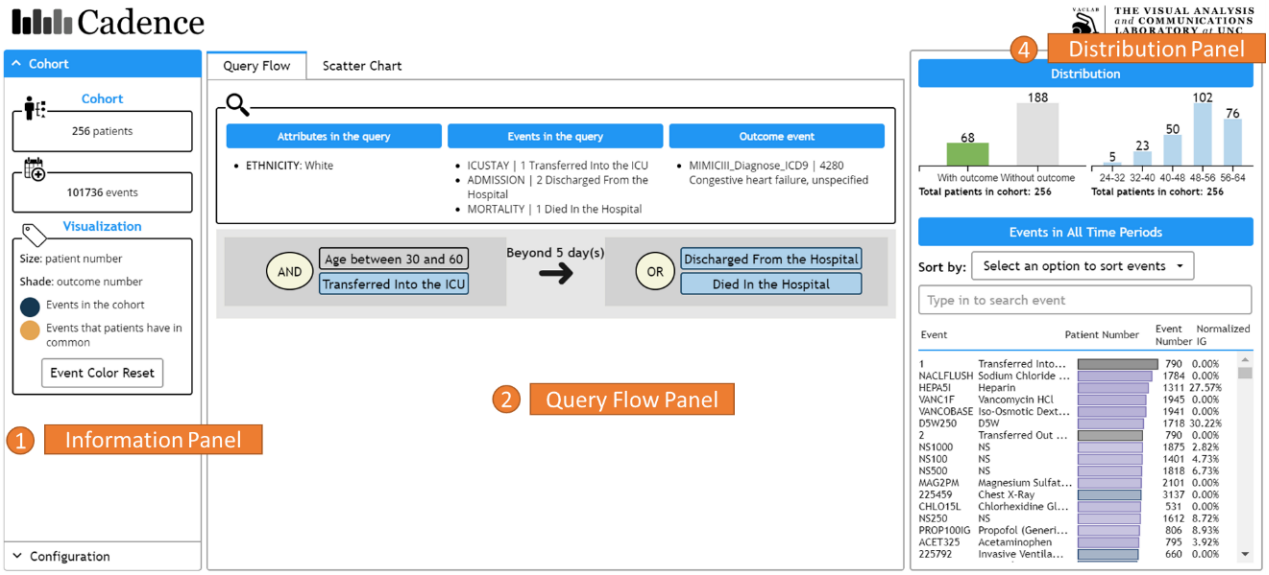


Figure 1. The main interface contains 4 panels: (1) the Information Panel, (2) the Query Flow Panel, (3) the Scatter Chart Panel (shown in Figure 2) and (4) The Distribution Panel.

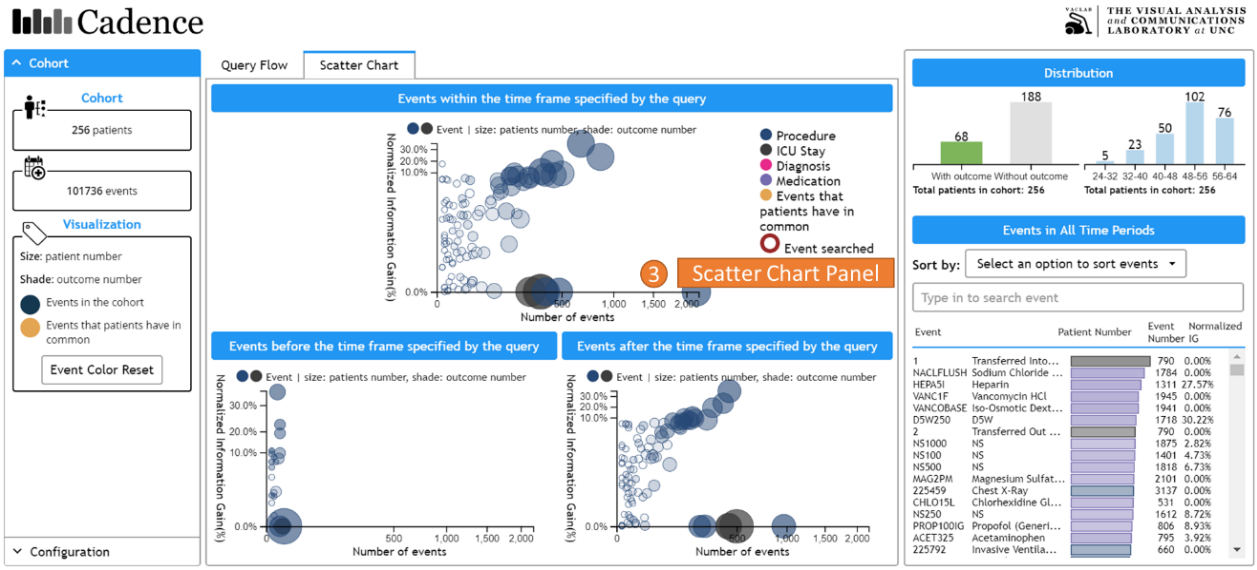


Figure 2. The Scatter Chart Panel (3) is located on a tab next to the Query Flow Panel.

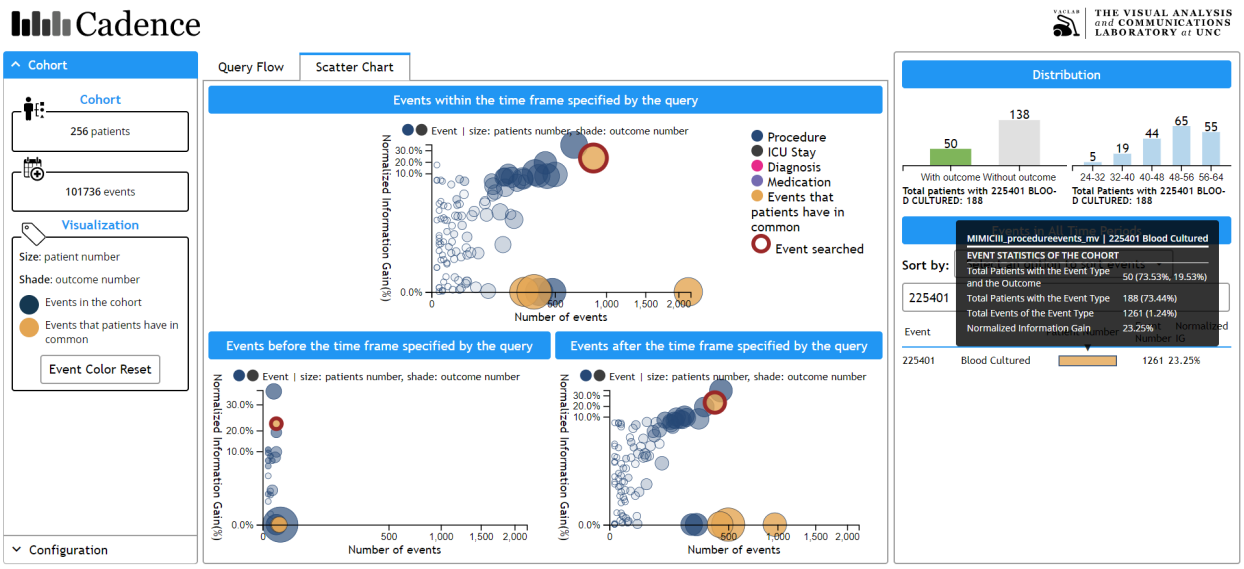


Figure 3. Users can search for specific event types via the distribution panel. In this example, the user types in 225401, the event code associated with "Blood Cultured." The corresponding circles are highlighted by a red circle in the 3 scatter charts (if the event is present).

query's temporal constraints (R1). The example shown in Figure 1 shows a query that returns 256 patients and 101,736 medical events. This view also contains a legend for the color coding used for highlighting throughout the interface.

Query Flow Panel. The query logic and criteria for both attributes (R1) and temporal constraints (R2) are shown in the query flow panel (Figure 1, panel 2). In addition, the query panel includes a definition of an outcome measure related to the research question for which the cohort is being defined. Instead of using forms and text to represent the query structure, this panel uses a flowchart-like representation designed for intuitive interpretation. The representation supports arbitrary levels of nesting and sequential constraints, while enabling users to read the query without any prerequisite knowledge of SQL or databases. For example, the query in Figure 1, panel 2 can be interpreted as "I want the patients who were transferred into the ICU between ages 30 and 60, at least 5 days before the patients were either discharged from the hospital or died."

The event boxes in the query flow panel can be clicked to trigger coordinated event highlighting in other panels. This enables users to see commonly co-occurring events, which can help inform query refinement (R3). For example, when the user clicks "Died in the Hospital" in Figure 4, all views in the interface will highlight other event types which frequently co-occur with hospital deaths. In this case, highlighted events include "Transferred into/out of the ICU" and "Chest X-Ray" (Figure 4).

Scatter Chart Panel. The scatter chart panel (Figure 2, panel 3) contains three scatter charts showing statistics for the event types that occur (1) before, (2) within, and (3) after the episode time frame defined by the temporal portion of the cohort query. In all scatterplots, bubbles represent individual event types and bubble size represents the number of patients with an event. The x-axis represents the number of occurrences for the event across all patients (each patient may have multiple occurrences of a single event type), while the y-axis represents a normalized Information Gain (IG) value. The IG value, based on a measure of entropy, measures how informative the occurrence of an event is with respect to a patient being included within (or excluded from) the query result set.

Intuitively, a high IG value indicates that an event can serve as a perfect predictor (positive or negative) for the inclusion (or exclusion) of patients within the query result. In contrast, an IG value of zero would mean that there is no association between the event occurring in a patient's history and that patient's inclusion (or exclusion) from the query result. The IG values are normalized to the overall frequency of an event within the entire patient database to enable comparison of scores across different events.

Finally, the shade of each bubble represents the average outcome for the patients who have the event. As in the query flow panel, all event bubbles can be clicked to trigger coordinated highlighting of co-occurring events.

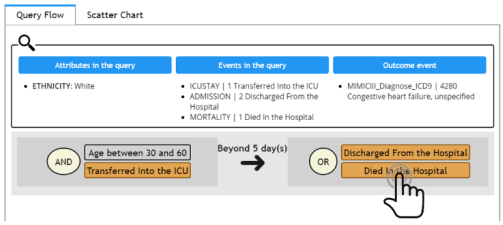


Figure 4. Event types can be clicked in any panel. In response, the system highlights events that co-occur with the clicked event. The highlights provide linked-coordinated views that help users understand relationships between events that may be informative for cohort refinement.

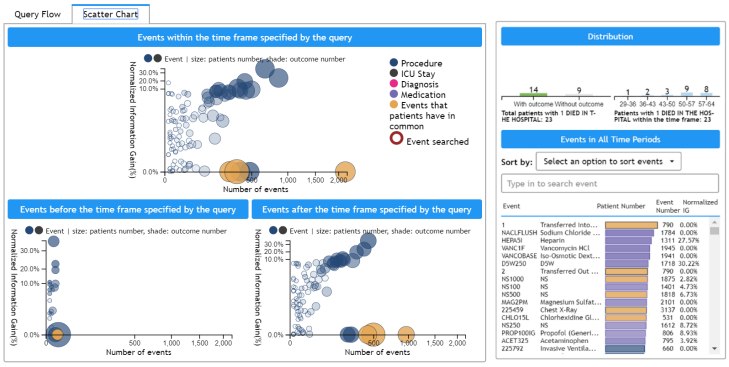


Figure 5. This figure shows highlights of co-occurring events resulting from the user clicking on the panel in Figure 4.

We note that it is common practice for visualizations to represent data returned by a query (e.g., the within scatter plot) (R1). However, our design also incorporates visualizations of data before and after the query time interval to provide contextual information to users which can help inform query refinement (R3).

Distribution Panel. The final panel in the interface (Figure 1, panel 4) shows additional details about the queried cohort (R1, R3). This includes age and outcome distributions, as well as a searchable histogram of event types. Users can enter a search in this panel to highlight a particular event type by event name or code. Searched events are highlighted in red as shown in Figure 3. As in other views, the individual event type can be clicked to trigger coordinated highlighting of co-occurring events.

3.2 Design Study with Users

The visual query system described above was motivated by challenges faced by current i2b2 users during cohort construction. Therefore, to evaluate the design, we recruited three current i2b2 users to take part in a design study.

Pre-Study Survey. To begin, a pre-study survey was sent to each of the three participants. This study gathered information about previous cohort query experiences with our institution's i2b2 system. Two of the three had used temporal constraints to define cohorts, but found it frustrating due to the slow and complicated processes of both the query construction/execution phase and the result exploration phase. The third user had not used this i2b2 feature at all. The pre-survey also asked users to describe the topics of their i2b2 query activity. This information was used to create initial queries in our prototype (which was not operationally deployed within the i2b2 environment) for subsequent study sessions. Both "simple" and "complex" initial queries were defined, enabling us to study the interface in different scenarios.

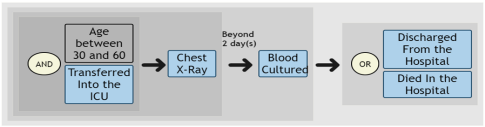
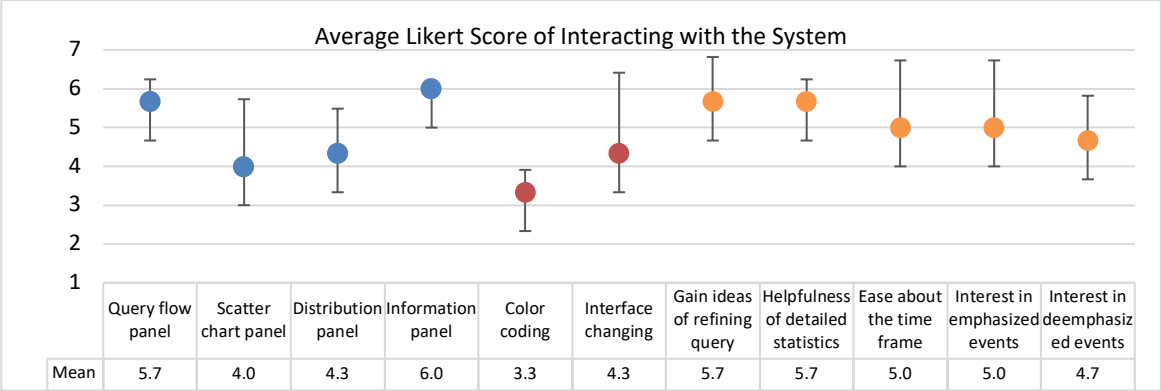


Figure 6. A complex query from the study.

Study Sessions. Each participant took part in a single study session lasting approximately 60 minutes. Each session began with a moderator providing an overview of the interface design and functionality. Participants were allowed to ask questions and interact with the system on their own. Then, after becoming familiar with the system, participants were asked to perform 2 groups of 8 tasks (a total of 16 tasks per participant): one group for a simple query and one for a complex query (e.g., Figure 6). The participants were asked think aloud while completing the tasks, and a screen recording application was used to record the entire session. The 8 tasks for each group were arranged from easy to hard, with tasks patterned after the iterative query refinement process typical of our target users. Task 1 was to interpret the meaning of the initial pre-defined query. Tasks 2-5 were focused on understanding the data returned by a query, including basic statistics about the cohort and related event types. Time-to-completion was recorded for tasks 1-5. Tasks 6-8 were open-ended to see how the system helped the user plan follow-up refinement queries performed using the user interface.

Post-Session Survey and Interview. After completing the study tasks, participants completed a post-session survey and sat for a semi-structured interview with the study moderator. The post-session survey included 7-point Likert scale subjective questions about each user's experience with our prototype system. The semi-structured interview provided an opportunity for the participants to provide deeper qualitative feedback, and for the moderator to probe the reasons behind observed user behaviors.



	Query flow panel	Scatter chart panel	Distribution panel	Information panel	Color coding	Interface changing	Gain ideas of refining query	Helpfulness of detailed statistics	Ease about the time frame	Interest in emphasized events	Interest in deemphasized events
Mean	5.7	4.0	4.3	6.0	3.3	4.3	5.7	5.7	5.0	5.0	4.7

Figure 7. Likert scores of the post-session survey. Blue circles represent the ease of interpreting different panels; red circles represent color coding and links between panels; orange circles represent utility for query refinement.

3.3 Results and Discussion

Overall, the participants found the visual interface effective and relatively easy to use. Participants were able to correctly answer the vast majority of factual questions (67 out of 72, or 93%, for tasks 1-5). Moreover, the questions were answered relatively quickly, completing all tasks within the time allotted to the study session. All three participants agreed (1 strongly agreed, 2 somewhat agreed) that the interface was easy to use for the purposes of query refinement. However, not all feedback was positive. We provide a thematic analysis of the study results below. The analysis highlights key user needs and goals, along with strengths and weaknesses of the proposed design.

Result 1: The size of the cohort was the primary concern when defining a query. Throughout the study, it was observed for all participants that they paid the most attention to the number of patients within a cohort rather than any of the other statistics provided by the interface. This finding reflects what users reported about their i2b2 usage, in which undesirable cohort size (too large or too small) was one of the key motivations behind the desire to perform query refinement.

Reflecting this, users found the variety and richness of additional data provided to be more than they desired. When reading the various measures associated with an event type, users found they had to filter through the "unnecessary" information to find what was must useful. One of the participants pointed out that the size of the bubble (representing the number of the patients) was the most important and helpful aspect to look at when exploring the data because it could give a direct indication of the expected refined cohort size. The statistics related to outcome were found to be less valuable. This suggests that users were less interested in changes to cohort composition (e.g., due to confounding relationships between events) than simple size of cohort. We believe that this reflects the current work pattern of defining a cohort as a first step before embarking on any outcome analysis activities. However, in other work from our lab, we have shown that earlier attention to outcome has the potential help yield higher quality cohorts.[22,23] Future work should explore how to effectively integrate outcome analysis tasks earlier into the process.

Result 2: The time frame concept was helpful for the temporal query, but better filtering is needed. Tasks 3-5 required users to answer questions related to the time frame concept depicted via the three scatter charts. Users performed these tasks more slowly than the more traditional non-temporal tasks (tasks 1-2). We believe that this was in part due to the more complex nature of the questions. However, we also observed that because the interface provided multiple paths to answer the questions, users spent time to confirm answers through multiple mechanisms. This is not necessarily a problem, but did appear to contribute to slower task completion times. The challenge was made more pronounced by difficulties in interpreting the color-coding used to link co-occurring events across scatter charts and the distribution panel. These difficulties, we believe, are reflected in the lower Likert scores associated with the color coding and the interactions for changing between panels in Figure 7.

Despite these difficulties, users found the time frame concept (before, within, after) a useful way to examine a query result for refinement. A typical use case, for instance, would be to locate an event type to see if it exists before, within, or after the query interval to determine what would happen to the cohort if that event type were added as an additional constraint in a follow-up query. The participants all understood the "big picture" for why this information was helpful. However, it required effort to sift through the various event types and integrate the various pieces of information located across the multiple panels. This is reflected in the performance statistics for task 5, which required users to click, hover, and search in combination to answer questions related to the events at various points in the time frame.

Task 5 was one of the most complex tasks, and not surprisingly had the most wrong answers. One challenge we observed was that information that required scrolling (e.g., because of a long histogram in the distribution panel) was often overlooked. This points to the challenge associated with presenting users with large numbers of variables.

Nonetheless, users expressed appreciation for the value provided by this complex feature. For instance, one participant mentioned that his/her research often requires restricting a cohort to the clinical events in a specific admission period. Using their typical tools, the participant had little confidence that the patients returned by a query were fully qualified to be members of the cohort. Often they discovered eligibility problems only after initially recruiting that patient to join a study. They felt the time concept would enable early discovery of this type of information. Another participant found that the system helped uncover specific events that had unexpected impacts on cohort size. For example, the participant mentioned a story about obtaining a very small cohort from an initial query, and that their typical tools have no way of showing that—in actuality—a small change in temporal requirements might result in a larger cohort size. With existing i2b2 tools, the participant would have to use trial and error to blindly search for ways to loosen their constraints. In contrast, the time frame concept would enable users to see this impact right away.

Result 3: The query flow was helpful for visualizing the query logic and was intuitive to understand and edit. All three participants correctly interpreted the meaning of the pre-defined initial query using the query flow panel, and all three felt that it was relatively easy to understand. It was also observed that follow-up queries were intuitively represented in a similar manner. One of the participants contrasted the view with the i2b2 interface provided by our institution's query system, and suggested that the query flow view gave them more confidence regarding how to define a query to meet their needs.

Another interesting suggestion by one of the participants had to do with query evolution during the cohort refinement process. Their current approach using i2b2 is to define a series of independent named queries. However, this approach is unwieldy to maintain and doesn't capture how cohorts relate to one another. The participant suggested that it would be helpful to use a view like the query flow to capture how a query definition changes over time as it is refined.

Result 4: Three factors highly impacted the follow-up queries: domain expertise, expected cohort size, and IG. When refining a query, users found the scatter chart panel and the distribution panel most useful. They way these panels were used, and the types of follow-up queries that were developed, were influenced by three key factors: (1) domain expertise, (2) the expected size of the cohort, and (3) the normalized IG value.

Domain expertise. The initial queries were closely related to the day-to-day research areas of participants 1 and 2, but more tangential to participant 3. This familiarity with the research question had a major impact on user behavior. It was relatively easy for participants 1 and 2 to internalize the data and quickly plan future actions. In contrast, participant 3 had a more difficult time understanding what he/she was looking for within the data.

The expected size of the cohort. It was observed that all participants tended to focus on cohort size, and mentally focused on aspects of the visualization that conveyed size-related information. This meant most attention was focused on bubble sizes in the scatter charts and on top events in the event histogram of the distribution panel.

Information gain (IG). As a secondary factor after cohort size, users did examine the IG metric. Most meaningfully, participants tended to scan the events from the top of the scatter chart (representing events with the highest IG). Users felt it was less valuable in part because it was difficult for them to interpret. We hypothesize that if this added information were to become more familiar, it might be leveraged more aggressively to guide the refinement process. Longer term user studies will be required to better understand its value.

4 Conclusion

The paper described the preliminary design and prototype for a visual query system designed to help users select and refine a cohort to be used in a retrospective study. The proposed design aims to address aspects of two difficult challenges which make iterative query refinement complex and time consuming for researchers: (1) the high-dimensionality of medical data, and (2) the temporal nature of many queries (e.g., "patients with a specific medical procedure within X days after diagnosis"). The paper enumerated a set of design requirements, described a set of visualization-based interface designs to support query construction and refinement, and presented results from a design study conducted to evaluate the proposed designs. A thematic analysis of the study results shows both strengths of the visual query design and opportunities for future improvements.

5 Acknowledgements

This project supported in part by the National Science Foundation under Grant No. 1704018.

References

1. Glenn, N. *Cohort Analysis*. (SAGE Publications, Inc., 2005). doi:10.4135/9781412983662
2. Murphy, S. N. *et al.* Combining clinical and genomics queries using i2b2 – Three methods. *PLoS ONE* **12,** (2017).
3. Hripcsak, G. *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud. Health Technol. Inform.* **216,** 574–578 (2015).
4. Krause, J., Perer, A. & Stavropoulos, H. Supporting Iterative Cohort Construction with Visual Temporal Queries. *IEEE Trans. Vis. Comput. Graph.* **22,** 91–100 (2016).
5. Klimov, D., Shahar, Y. & Taieb-Maimon, M. Intelligent visualization and exploration of time-oriented data of multiple patients. *Artif. Intell. Med.* **49,** 11–31 (2010).
6. Catarci, T., Costabile, M. F., Levialdi, S. & Batini, C. Visual Query Systems for Databases: A Survey. *J. Vis. Lang. Comput.* **8,** 215–260 (1997).
7. Hibino, S. & Rundensteiner, E. A. User Interface Evaluation of a Direct Manipulation Temporal Visual Query Language. in *Proceedings of the Fifth ACM International Conference on Multimedia* 99–107 (ACM, 1997). doi:10.1145/266180.266342
8. Gotz, D. & Stavropoulos, H. DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data. *IEEE Trans. Vis. Comput. Graph.* **20,** 1783–1792 (2014).
9. Jin, J. & Szekely, P. Interactive Querying of Temporal Data using a Comic Strip Metaphor. in *2010 IEEE Symposium on Visual Analytics Science and Technology* 163–170 (2010). doi:10.1109/VAST.2010.5652890
10. Aigner, W. & Miksch, S. CareVis: Integrated visualization of computerized protocols and temporal patient data. *Artif. Intell. Med.* **37,** 203–218 (2006).
11. Bade, R., Schlechtweg, S. & Miksch, S. Connecting Time-oriented Data and Information to a Coherent Interactive Visualization. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 105–112 (ACM, 2004). doi:10.1145/985692.985706
12. Wongsuphasawat, K., Plaisant, C., Taieb-Maimon, M. & Shneiderman, B. Querying Event Sequences by Exact Match or Similarity Search: Design and Empirical Evaluation. *Interact. Comput.* **24,** 55–68 (2012).
13. Burch, M., Beck, F. & Diehl, S. Timeline Trees: Visualizing Sequences of Transactions in Information Hierarchies. in *Proceedings of the Working Conference on Advanced Visual Interfaces* 75–82 (ACM, 2008). doi:10.1145/1385569.1385584
14. Du, F., Plaisant, C., Spring, N. & Shneiderman, B. EventAction: Visual analytics for temporal event sequence recommendation. in *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)* 61–70 (2016). doi:10.1109/VAST.2016.7883512
15. Cibulski, L. *et al.* ITEA—Interactive Trajectories and Events Analysis: Exploring Sequences of Spatio-temporal Events in Movement Data. *Vis. Comput.* **32,** 847–857 (2016).
16. Monroe, M., Lan, R., Lee, H., Plaisant, C. & Shneiderman, B. Temporal Event Sequence Simplification. *IEEE Trans. Vis. Comput. Graph.* **19,** 2227–2236 (2013).
17. Wang, T. D. *et al.* Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. *Vis. Comput. Graph. IEEE Trans. On* **15,** 1049–1056 (2009).
18. Wang, T. D. *et al.* Aligning temporal data by sentinel events: discovering patterns in electronic health records. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 457–466 (ACM, 2008). doi:10.1145/1357054.1357129
19. Monroe, M. *et al.* The challenges of specifying intervals and absences in temporal queries: a graphical language approach. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 2349–2358 (ACM, 2013). doi:10.1145/2470654.2481325
20. Wongsuphasawat, K. & Gotz, D. Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization. *IEEE Trans. Vis. Comput. Graph.* **18,** 2659–2668 (2012).
21. Partl, C. *et al.* enRoute: Dynamic Path Extraction from Biological Pathway Maps for In-depth Experimental Data Analysis. in *2012 IEEE Symposium on Biological Data Visualization (BioVis)* 107–114 (2012). doi:10.1109/BioVis.2012.6378600
22. Gotz, D., Sun, S. & Cao, N. Adaptive contextualization: Combating Bias During High-Dimensional Visualization and Data Selection. in *Proceedings of the 21st International Conference on Intelligent User Interfaces* 85–95 (ACM, 2016). doi:10.1145/2856767.2856779
23. Gotz, D., Sun, S., Cao, N., Kundu, R. & Meyer, A.-M. Adaptive Contextualization Methods for Combating Selection Bias during High-Dimensional Visualization. *ACM Trans. Interact. Intell. Syst.* **7,** 1–23 (2017).