Visual Assessment of Cohort Divergence During Iterative Cohort Selection

David Gotz
Carolina Health Informatics Program
School of Information and Library Science
University of North Carolina at Chapel Hill
216 Lenior Drive; CB 3360
Chapel Hill, NC 27599
gotz@unc.edu

Shun Sun
School of Information and Library Science
University of North Carolina at Chapel Hill
216 Lenior Drive; CB 3360
Chapel Hill, NC 27599
shunsun@live.unc.edu

ABSTRACT

Large-scale repositories of secondary-use patient data are emerging as a critical resource for both clinical and epidemiological research. Motivated by this opportunity, a variety of interactive visual analysis methods have been developed to make the use of this data more efficient and accessible. These techniques often combine interactive filters and on-demand computational analysis to allow ad hoc cohort exploration and refinement. This approach has indeed made it possible to quickly select and revise cohorts during analysis. However, the seemingly simple filters supported by these tools can produce dramatic—and often unseen—confounding effects on the makeup of the cohort across the thousands of variables often found in real-world medical data. This poster presents an approach to measuring and visually conveying to users the degree of drift in representation during iterative visual cohort selection.

CCS Concepts

•Human-centered computing → Visualization; Information visualization; •Applied computing → Health informatics;

1. INTRODUCTION

Interactive visual analysis systems support a highly dynamic and iterative cohort selection process. Using direct manipulation methods, analysts can quickly apply complex sequences of filters to a dataset to identify patients of interest for additional analysis. Such techniques have been applied successfully in both the visualization literature (e.g., [4]) and in the health informatics community (e.g., [2]).

Such methods can make it relatively easy for investigators to rapidly identify sets of patients based on specific inclusion and exclusion criteria. However, users typically apply filters to only a very tiny fraction of the thousands of variables found in modern electronic health record systems. Given the significant confounding relationships between these many variables, coupled with the quick speed with which additional filters can be applied, investigators are at great risk of producing cohorts with large, unseen, and unexpected selection biases.

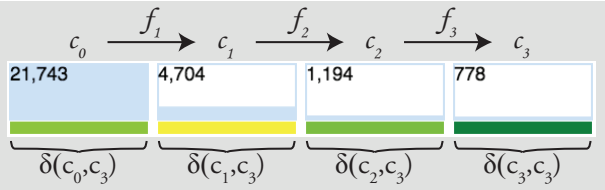

Figure 1: A glyph-based visualization depicts the chain of cohorts defined by a user as filters are iteratively applied. The rectangular glyphs are ordered sequentially from left to right. Each cohort's glyph depicts both its size and its divergence measure when compared against the user's current cohort.

To help users understand the unseen changes in cohort makeup over the course of an iterative selection process, we have developed tools for visual assessment of cohort divergence. These tools, which measure and convey how representative one cohort is of another, have been integrated into our prototype visual analytics system, Tempo Analytics. This prototype, shown in Figure 2, adopts visualization and interaction designs that are inspired by our previous work on temporal cohort visualization [1].

2. METHODS

Users of our system begin by defining an initial patient cohort (c_0) via a visual query interface. The query results are visualized, after which users can select additional filtering criteria (f_1) to define a more focused cohort (c_1). This process repeats until the user is satisfied with the final cohort (c_n). We model this iterative process as a *cohort chain*, which we note as $\{c_0, c_1, \ldots c_n\}$.

For each pairwise combination of cohorts, we compute a divergence metric $\delta(c_i, c_j)$ that quantifies the similarity in patient populations between the two patient groups. Our measure builds upon the Hellinger distance [3], a metric that represents the similarity between two probability distributions. We compute this distance individually for each

VAHC '15 October 25, 2015, Chicago, IL, USA

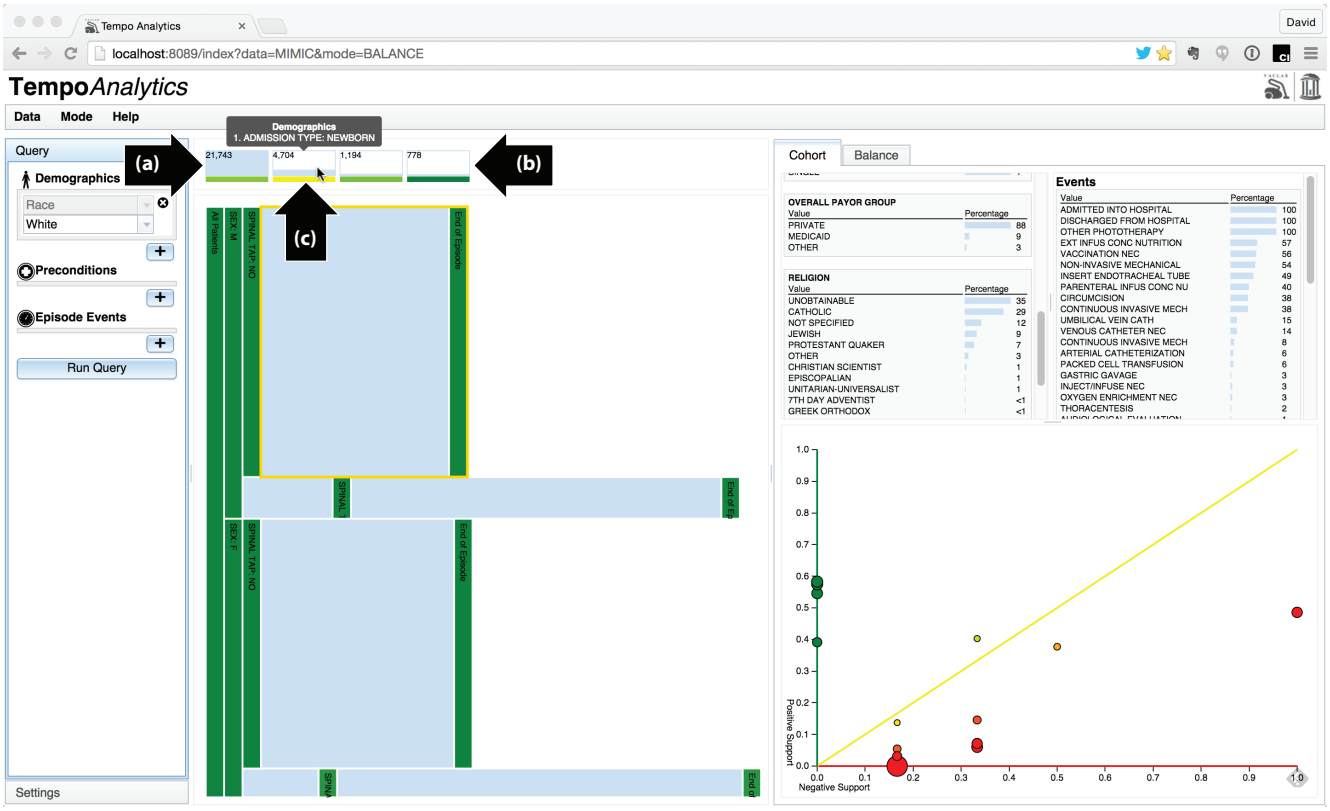

Figure 2: This screen capture shows an analysis with four cohort refinement steps. These steps narrow the population from (a) 21,743 to (b) 778 patients. The final cohort is least similar to the population at (c) step 2 as indicated by the yellow bar underneath the cohort marker. The second cohort of 4,704 patients was created by filtering to include only newborn admissions, a unique population which resulted in many confounding changes in variable distributions. Subsequent filtering steps narrowed the analytic focus to a newborn population with additional medical complications. This produced a newborn cohort with variable distributions that were more representative of the overall population by excluding the unique "healthy newborn" patients.

variable, then average the univariate measures into a single aggregate score for the cohort pair. A value of zero indicates two cohorts whose variables have identical distributions. Increasing metric values reflect increasingly diverging distributions.

Our prototype illustrates this data—both the cohort chain and associated statistics—using a glyph-based visual representation. Shown in Figure 1, the chain is depicted as a series of rectangular glyphs arranged sequentially from left to right. Each cohort c_i has its own glyph, with new glyphs added interactively as new filters are applied.

Each glyph has two components, with the top section showing cohort size. The lower section, meanwhile, represents the divergence measure comparing the glyph's cohort with the final cohort in the chain (c_n). The measure's value encoded with color using (by default) a green-yellow-red color gradient. Solid green represents a value of zero, while high values (large differences between cohorts) represented with red. The colors for each cohort are updated dynamically as new cohorts are added to the chain and new distance measures are computed.

As shown in Figure 2, the cohort chain visualization is placed prominently within our user interface for cohort selection. By placing it front and center within the display, users are immediately provided with a indicator of the selection bias introduced by confounding relationships as they emerge during the cohort selection and refinement process.

3. ACKNOWLEDGMENTS

This work was made possible in part by a Data Fellow award from the National Consortium for Data Science (NCDS).

4. REFERENCES

[1] D. Gotz and H. Stavropoulos. DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1783–1792, 2014.

[2] S. N. Murphy, V. Gainer, and H. C. Chueh. A Visual Interface Designed for Novice Users to find Research Patient Cohorts in a Large Biomedical Database. *AMIA Annual Symposium Proceedings*, 2003:489–493, 2003.

[3] D. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2002.

[4] Z. Zhang, D. Gotz, and A. Perer. Iterative cohort analysis and exploration. *Information Visualization*, OnlineFirst, Mar. 2014.