Visualization of EHR and Health Related Data for Information Discovery

Vivian West[1] David Borland[2] W. Ed Hammond[1]
[1]Duke Center for Health Informatics, Duke University
[2] Renaissance Computing Institute, The University of North Carolina at Chapel Hill

Abstract

In this paper we describe research we are conducting in response to a Program Announcement solicited by the Assistant Secretary of Defense for Health Affairs, Defense Health Program. The amount of information in Electronic Health Record (EHR) systems is growing rapidly with the inclusion of disparate forms of data from a number of new sources, i.e. genomics and imaging data. EHR systems will continue to grow as more healthcare data is digitized. As data in EHRs grows, there is increasing interest in understanding what information and knowledge these large data sets represent.

Data visualization techniques offer an opportunity to explore and understand large data through novel approaches. Our research seeks to visualize health care data from electronic health records (EHR) and other health related data. Our approach is informed by retrospective data queries using DEDUCE, a query tool developed at Duke University.

Keywords: Electronic health records, health related data, information visualization

Introduction

Visualization of genomic data is used to understand data structures. Geospatial applications have revealed patterns related to risk factors in environmental health,[1,2] and visualization methods of limited data sets have been used for clinical decision support.[3,4] Data from EHRs and other health related data, however, are displayed primarily through techniques that have been used for many years, e.g. fishbone diagrams for lab values, or by using charts and graphs. There have been few successful attempts to visualize massive amounts of disparate health care data.

Effective visualization techniques of large health data sets will allow users to see patterns they would not otherwise see. With many sources of health related data containing many parameters, the ability to visually explore the collective data has the potential to reveal valuable information.[5] There are many data elements and attributes in healthcare data. We propose that grouping and aggregating related data elements via a priori categorization (e.g. laboratory results or vital sign data) or data-driven methods (e.g. correlation) will facilitate developing visualization techniques that will allow users to see patterns in large data and elicit further inquiry of the data. We also believe the user should be able to further explore the data by opening the visual representation of a set of data elements to see trends representing aggregated data and drilling down even further to the subsets of the data. By having an interactive visualization, the ability to explore and gain a deeper understanding[6] of what the data represent will encourage adoption of the visualization technique, assuming the visual presentation minimizes cognitive burden.

Related Work

There are numerous reports in the literature related to data visualization in health care, most focusing on the technical aspects of visualization, medical imaging, and genomics. A number of prototypes have been also been reported. LifeLines, first described in 1996 by Plaisant and colleagues,[7,8] was used to visualize health data across a personal health record using timelines. Lifelines evolved to become Lifelines2, a visualization tool using categorical point event data across multiple records. More recently, Eventflow, similar to Lifelines2, also addresses the need to have a system to support interval events.[9]

Novel visualization techniques using EHRs was somewhat limited until 2009 when the HITECH Act mandated EHR implementation. In addition to evolving changes to LifeLines, several prototypes are in various stages of development. Most reported techniques are interactive, allowing the user to explore data incorporated as one visual display. For example, Zhang, et. al.[10] use a radial starburst visualization of multiple data points from one health record permitting users to drill down on data to single time points. Klimov and Shahar describe a prototype called VISITORS (Visualization of Time-Oriented Records) using time-oriented data sets with an interface to explore longitudinal values.[11] These approaches are similar to that we are taking, but we believe the historical queries and identification of the data elements and clusters will enhance visualization of relevant data.

Methods

Using historical data queries of Duke's EHR system (called DEDUCE) we will identify what data elements are in queries and classify them according to the types of information sought (e.g. outcomes, outliers of treatment

methods). Groups of related data elements will be incorporated into a visualization that allows a quick comparison of the data from a large population with the ability to view trends over time within a chosen measure.

The following example illustrates our approach using the Department of Defense mock EHR data. We will look at aggregated health related data from an Army unit pre- deployment using visualization to discover differences within the group. We will then compare the same data elements post-deployment to identify changes. These time periods can be compared with the group later diagnosed with post-traumatic stress disorder to identify outliers and what data elements might have caused the outlier. In this example, Army personnel between 25-30 years old who have been deployed can be compared to the population of all adults in the system, or all Army or Marine or Navy personnel. We will employ visualization methods that show aggregated groups of data elements with a distribution per population, with the ability to drill down in the data and display longitudinal data for selected data elements.

The key to selecting the most effective method of visualization is to understand how to address the informational value of the data. We expect classes of data elements with the greatest variation to stand out. We will statistically pre-process data as an enhancement to visualization, eliminating null associations and unimportant variables (statistically). In comparing groups, the visualization method should clearly show differences. Further examination of data should also permit the easy application of different filters and the ability to hone down on subsets of data.

Discussion

1. Radial Coordinates Visualization. We have developed an initial multivariate visualization tool in the statistical programming environment R, using the RGL package to enable real-time interactive visualizations. This radial-coordinates visualization prototype is inspired by parallel-coordinate[12,13] and star plot[14] multivariate visualization techniques.

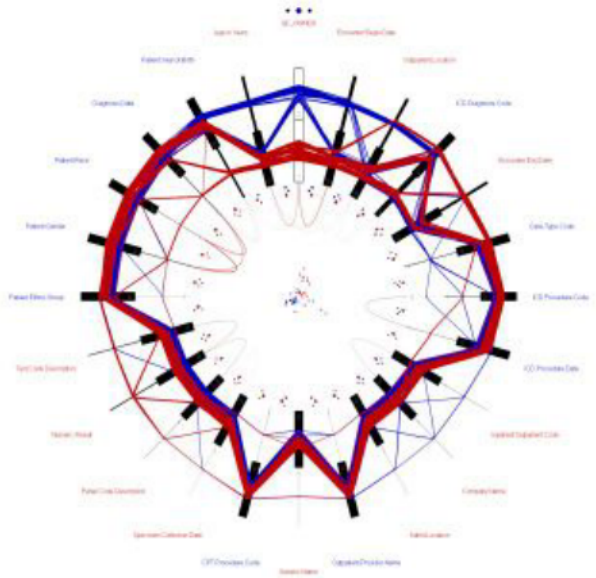

Figure 1: Radial coordinates visualization of DEDUCE queries.

Figure 1 shows an example radial coordinates visualization using queries from the top two users in Duke's DEDUCE EHR query tool. Each line represents a query, and the value for each axis represents how often that data element was used in the given query (typically zero or one). The lines are colored by system user. A circular layout of the axes has the advantage of a square aspect ratio when compared to standard parallel coordinates axes, which can be beneficial for large numbers of axes. Within this framework we have looked at additional improvements to standard parallel coordinates techniques, such as showing data distributions directly for each axis based on data type. For continuous data we display a box-and-whiskers plot (not shown in Figure 1), for discrete integer-valued data we display a histogram with bin width proportional to number of entities with that value, and for categorical data we display a stacked bar chart, with bar length proportional to number of entities with that value. This enables rapid evaluation of the various data types for a heterogeneous dataset, and of the distribution for each variable. In addition, we introduce line spreading to mitigate the problems of multiple lines collapsing to a single data point for discrete and categorical data, extending the parallel sets method (http://eagereyes.org/parallel-sets) to enable visualization of individual data entities, and the incorporation of non-categorical data.

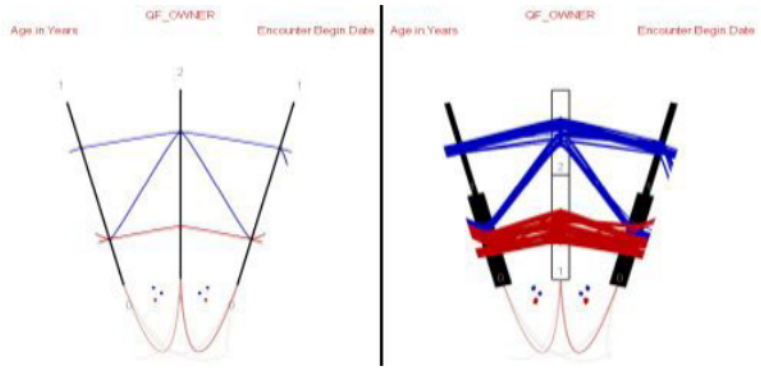

Figure 2: Line spreading (right) for discrete and categorical data enables improved visualization of multiple entities with the same value.

The close-up in Figure 2 illustrates the improvement possible using axis visualization with line spreading (right). The "QF_OWNER" categorical data (user 1 vs. user 2) is displayed using a stacked bar chart, with segment length proportional to number of entities with that value (there are slightly more queries from user 1), and individual lines are spread out within each bar segment based on their position on neighboring axes. The other two discrete integer-valued data elements are displayed using a histogram with bin width proportional to number of data entities with that value, and individual lines are spread out within each bin. With the visualization on the right it is much easier to follow individual lines between axes, and to see clustering of lines.

Axis-ordering is a well-known problem with parallel-coordinates techniques. We have experimented with a number of techniques for clustering axes based on correlation between axes. We also utilize correlation to flip axes to try to minimize line crossings, based on positive or negative correlation with a given axis. To enhance these techniques we also draw curved arcs connecting axis pairs, with opacity and line width proportional to correlation magnitude, and color based on correlation polarity (blue = negative, red = positive). Colored axis labels indicate whether the axis has been flipped (blue) or not (red).

The central space in the radial coordinates visualization enables the display of supplemental visualizations. Inspired by Holten and vanWijk,[15] we draw pair-wise scatterplots just below neighboring axes, and in the center we draw a scatterplot of the first two principal components. In the future, we plan to enable a number of different visualizations to be placed here, chosen interactively by the user. Each scatterplot and the radial coordinates visualization are linked together, such that selection in any visualization is reflected in the other visualizations.

2. Force-Directed Network Visualization. We have also developed a force-directed node-and-link network visualization to investigate queries from DEDUCE queries, implemented in the Processing programming environment. Figure 3 shows the same data as Figure 1, with individual query data elements drawn as circles, and de-identified system users (in this case the top two users of the system) as squares

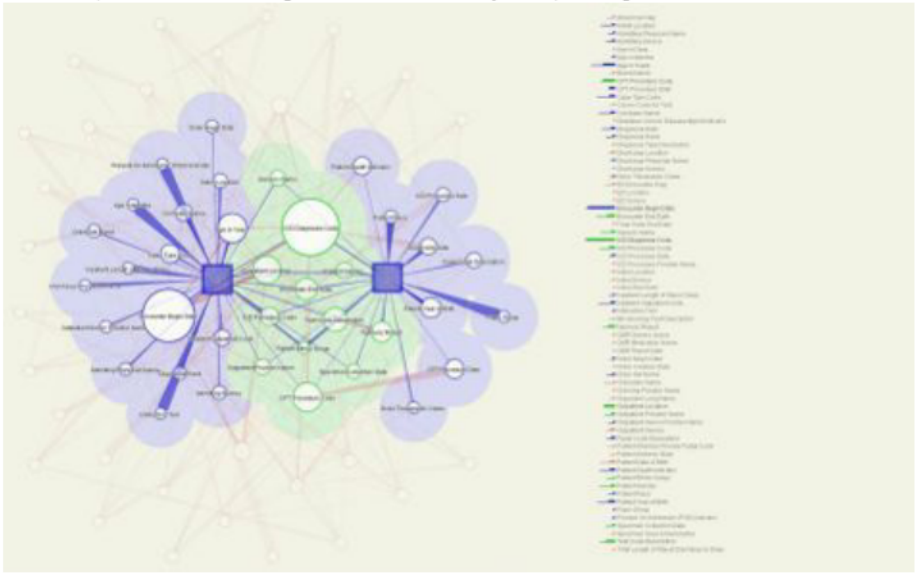

Figure 3: Force-directed layout visualization of DEDUCE queries.

The size of each circle represents how often it was used as a query element across all queries, and the size of each square represents the number of queries made by that user. Links between circles represent how often each element was used together in a series of queries, with each end scaled based on the relative importance at each end of the link. Links between circles and squares represent how often each user made a query on each element. Nodes are placed via a force-directed layout based on the overall strength of each link. In this example the user has highlighted the two users, which in turn highlights nodes connected to those users, while deemphasizing all other nodes. Nodes that are connected to both users are highlighted in green, whereas nodes that are connected to just one user are highlighted in blue. A full list of data elements is shown to the right, with horizontal lines representing the number of times each element was used across all queries (equivalent to circle size), and the number of other elements connected to. The user can interactively select nodes via the node-link diagram or the list of elements.

Some relationships are more easily discernible in one representation vs. the other. E.g. it is perhaps more readily apparent in Figure 3 that ICD Diagnosis Code is the most-used query element, and both users used that element, whereas in Figure 2 it is more apparent that Patient Gender, Patient Race, and Patient Diagnosis Date are all strongly correlated (i.e. they tended to be used together in queries), and that one of the users (red) included those elements more than the other. Our approach going forward will therefore combine such visualizations to enable multiple linked views of the data.

Conclusions

Compressing petabytes of health care data representing many data elements into various groups of related data presented visually with an interface that allows the user to interactively explore the data elements, to our knowledge never been done. There is the potential to detect causal relationships between various sets of data, which may lead to improved health care costs.

Acknowledgements

This work is supported by research funds from the Department of Defense, Award number W81XWH-13-1-0061.

References

1. Miranda ML, Edwards SE. Use of spatial analysis to support environmental health research and practice. NC Med J 2011;72:132-5.
2. Miranda ML, Edwards SE, Anthopolos R, Dolinsky DH, Kemper AR. The Built Environment and Childhood Obesity in Durham, North Carolina. Clin Pediatr (Phila) 2012.
3. Mane KK, Bizon C, Owen P, Gersing K, Mostafa J, Schmitt C. Patient Electronic Health Data–Driven Approach to Clinical Decision Support. Clinical and Translational Science 2011;4:369-71.
4. Mane KK, Bizon, C, Owne, P, Mostafa, J, Gersing, K and Schmitt, C. A Paradigm Shift: Electronic Health Records Data in Clinical Practice (Abstract). In: 2011 CTSA Annual Informatics Meeting. Natcher Conference Center, NIH Campus, Bethesda, MD; 2011:64-5.
5. Gershon N, Eick SG. Visualization's new tack: Making sense of information. Spectrum, IEEE 1995;32:38-40, 2, 4-7, 55-6.
6. Shahar Y, Goren-Bar D, Boaz D, Tahan G. Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. Artif Intell Med 2006;38:115-35.
7. Plaisant C Milash B, Rose A, Widoff S, Shneiderman B. LifeLines: Visualizing Personal Histories. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1996:221-227.
8. Plaisant C, Mushlin R, Snyder A, Li J, Heller D, Shneiderman B. LifeLines: using visualization to enhance navigation and analysis of patient records. *Proc. AMIA Symp*. 1998:76-80.
9. Lifelines2: Discovering Temporal Categorical Patterns Across Multiple Records. http://www.cs.umd.edu/hcil/lifelines2/. Accessed September 5, 2013.
10. Zhang Z, Wang B, Ahmed F, et al. The Five W's for Information Visualization with Application to Healthcare Informatics. *IEEE transactions on visualization and computer graphics*. Jun 3 2013.
11. Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. *Artif. Intell. Med*. May 2010;49(1):11-31.
12. d'Ocagne M. Coordonnees parallels et axiale. Gautier-Villars, Paris 1885.
13. Inselberg A. The plane with parallel coordinates. *The Visual Computer*. 1(2):69-91.
14. Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth, 1983.
15. Holten and vanWijk, Evaluation of Cluster Identification Performance for Different PCP Variants, Computer Graphics Forum, 29(3), 793-802, 2010.