AD________________

Award Number:
W81XWH-13-1-0061

TITLE:
Novel Visualization of Large Health Related Data Sets

PRINCIPAL INVESTIGATOR:
William Ed Hammond, PhD

CONTRACTING ORGANIZATION:
Duke University
Durham, NC 27705-4677

REPORT DATE:
March 2014

TYPE OF REPORT:
Annual Report

PREPARED FOR:
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:
Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*
March 2014

2. REPORT TYPE
Annual Report

3. DATES COVERED *(From - To)*
25 Feb 2013 to 24 Feb 2014

4. TITLE AND SUBTITLE

Novel Visualization of Large Health Related Data Sets

5a. CONTRACT NUMBER
W81XWH-13-1-0061

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Hammond, William, E; West, Vivian; Borland, David; Akushevich, Igor; Heinz, Eugenia, McPeck
email: william.hammond@dm.duke.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Duke University
2200 W. Main St, Ste 710
Durham, NC 27705-4677

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

US Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT Using retrospective data queries to understand what information clinicians seek from health care data, we have identified data elements and are looking at combinations of data elements used in queries. We are developing various visualization techniques that can be used to present the informational content in large databases, expecting that visualization of this data will present or "discover" information without specific hypotheses. Groups of related data elements will be incorporated into a novel visualization that allows a quick comparison of the data from a large population, with the ability to view trends over time within a chosen category. We are exploring the ability to compress 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 understand big data from the perspective of the entire military, different branches of service, military ranks and job specialties, ages, and geographical deployment areas. There is the potential to detect causal relationships between various sets of data, which may lead to improved health care and resiliency in military personnel, assist the DoD in strategic decisions related to personnel, and save millions of dollars in health care costs.

15. SUBJECT TERMS
Visualization, health care data, big data

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					USAMRMC
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	38	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

1. INTRODUCTION

With the growth of Electronic Health Record (EHR) data and other related healthcare databases, there is a need to understand what information and knowledge the data represent. Visualization offers an opportunity to explore and understand large data in unique and novel ways, permitting one to view data without the bias of an a priori decision of what is important. We hypothesize that data visualization is more effective than traditional methods of data exploration, and that the type of visualization is highly dependent on the types of data and nature of the queries and what someone is trying to learn from the data. The aims of this project are to: (1) use retrospective data queries to understand what information clinicians seek from health care data, identify what data elements and mixtures of data classes (laboratory data, demographic data, problems, therapies, physical examination data, or imaging data) are used in queries and what methods are used to analyze query results, and (2) create a matrix of data visualization methods used with specific data elements from multiple classes and test visualization of mixed data classes.

2. BODY

The timeline for completing our project milestones and associated tasks is behind several months. Although we had Duke IRB approval to begin the study in February 2013, we did not receive the Human Research Protection Office (HRPO) approval to begin working with actual clinical data until the end of September 2013, putting our task completion behind schedule. We adjusted our budget accordingly and anticipate requesting a no-cost extension to complete the project as planned using a slightly revised timeline.

While waiting for HRPO approval to use our clinical data, we worked with three different data sets: (1) automobile data that were included with R, the first statistical programming environment we used for visualizing data; (2) data of the various data elements queried in DEDUCE and the numbers of times each were queried; and (3) a population data set from Primary Care Trust (PCT) data from the National Health Services (NHS) representing longitudinal data from two to eight years for over 60 million individuals in the United Kingdom. Findings from our work with these data sets is detailed in 2.2.2 and 2.6.1 of this report.

Research accomplishments associated with the tasks within each of our nine milestones are detailed in 2.1 through 2.9. Section 2.10 describes future work to complete tasks and problems encountered.

2.1. Obtain access to queries of Duke's Clinical Data Warehouse

STATUS of Milestone: Completed.

At Duke we use an on-line query system called DEDUCE (Duke Enterprise Data Unified Content Explorer) to access data in the Decision Support Repository, which consists of hundreds of tables, some with hundreds of millions of rows of data, collected from over 4.4 million patients at Duke. DEDUCE was operationalized in 2008 and upgraded numerous times during the next three years. By 2011, it was recognized within the Duke medical community for its value in abstracting information from the Data Repository. Researchers can query numerous data elements and refine the query to facilitate exploration of aggregate clinical data in support of operations, quality, and research. Output from the queries is in the form of common-separated values (CSV) files, ASCII files, Excel files, or simple graphs.

Every query is saved on a Duke server. We wanted to obtain copies of retrospective queries to begin our project, but were delayed in accessing this query data until we received HRPO approval at the end of September 2013. We then requested retrospective query data from 01 Jan 2011 through 31 Jul 2013. We now have half of these queries, a significant amount of information for us to work with to complete this project. Copies of the queries have been transferred to a secure workspace on a firewall-protected Duke server for our research team to analyze.

2.2. Develop classification for queries

STATUS of Milestone: We will finalize our classification schema during Quarter 5.

2.2.1. Identify early use of data queries

To help us understand what information clinicians seek from data available to them and assist in our in-depth review of data queries, we conducted a survey of users of DEDUCE (Appendix A). A programmer in the Duke Office of Clinical Research used REDCap (Research Electronic Data Capture), a web-based application for building and managing online surveys to build and distribute the online survey, collect the results, and aggregate the responses in an Excel spreadsheet.

At the end of October 2013, an email request to complete the survey was sent to 482 users who within the last two years completed the required training course before being granted access to DEDUCE. A reminder email was sent on 18 Nov, and the survey was closed on 04 Dec 2013. A total of 61 people completed the survey, a 12.7% response rate. Respondents were asked to identify all of the reasons they ran queries. The most frequent aims for conducting queries were related to grant preparation, determining the prevalence of a particular population or getting other information for a grant (47.4%), quality improvement (46%), outcomes (33%), and to see if there were enough patients who would meet inclusion/exclusion criteria for participation in industry-sponsored clinical trials (28%). Of the additional clinical and non-clinical reasons for queries, 28% of the responses were also related to research activities.

When asked to approximate the number of times the respondents initiated queries in the last 2 years, the majority selected 1-4 (39%), with 29% selecting 15-19 queries (29.8%); 12% conducted queries more than 20 times. Respondents were asked to identify the types of information typically sought in a query, selecting all that applied. Results are noted in Table 1.

Query Data Requested	Percent(%)
Diagnoses	83
Demographics	75
Encounters	47
Procedures	46
Medications	39
Laboratory data	35
Physicians	28
Imaging data	23
Vital signs	16
Text for analysis	12
Geospatial data	9
Device information	5

Table 1. Percentage of users and type of data most frequently requesting using DEDUCE queries.

Respondents received the output from their queries as Excel tables (77%), ASCII files (18%), or CSV files (5%). Of those participating in the survey, 77% were satisfied always or most of the time with the information obtained as a result of the queries.

2.2.2. Based on the reason for the queries, group them accordingly.

Our preliminary findings using the early DEDUCE data set indicates that ICD9 codes are the most frequently queried data element, consistent with our survey data noted above. Using visualization techniques in R, we looked at the relationships between the data elements used in queries, how often each element was used together in a series of queries, and how often each user made a query on each. This process is further described in 2.4, and has helped us determine the most common relationships between the data elements used in the queries. (Refer to article in Appendix C for additional detail.). It is also useful as we develop our classification schema.

2.2.3. Obtain access to AHLTA de-identified data, and using work from the Duke queries and classes, compare for similarities and differences and revise classes as needed.

We will be unable to access AHLTA de-identified data, but do have a limited synthetic data set developed for the DoD. We have just obtained a copy of the synthetic data encoding tables and synthetic data schemas, which will help us better understand what some of the data is. We are preparing a request for a dataset of 1 million patients, which will be a useful test of the visualization techniques we have developed to date.

2.3. Identify data elements used in queries

STATUS of milestone: Identification of data elements is to be completed at the end of March 2014. Grouping into classes will be completed at the end of May 2014.

2.3.1. Identify which queries should be most meaningful to include in our analysis and the individual researchers associated with those queries.

We have completed a review of the queries to identify the data elements used and the frequency of their use.

We will next group the queries by individuals to evaluate if common data elements are used by specific types of researchers or those associated with specific clinical backgrounds, information that might be useful when we ask the respondents who agreed to work with us in identifying the most useful types of visualization, per the milestone in 2.4.

2.3.2. Group data elements into classes (e.g. laboratory data, demographics, medications).

A classification schema has been more difficult to develop than we had expected; there are many combinations that need to be considered. For example, while medications might be one category, a number of medications for adults and children differ, raising the issue of perhaps using a less inclusive classification, e.g. adult medication and pediatric medication. Within medications, we have also considered whether it is better to use drug classes to be more specific. Our review of the retrospective queries should give us more insight into an appropriate classification schema.

2.4. Explore alternative visualization methods of the data. Clinicians will use the Follow-up Questionnaire to compare alternate visualization of the data to the original presentation of the query data.

STATUS of milestone: Completion extended due to delayed HRPO approval to use clinical data.

We have used two different visualization techniques to explore query data from Duke's DEDUCE system. The first is a force-directed network visualization of query elements and system users developed using the Processing programming environment (Figure 1).

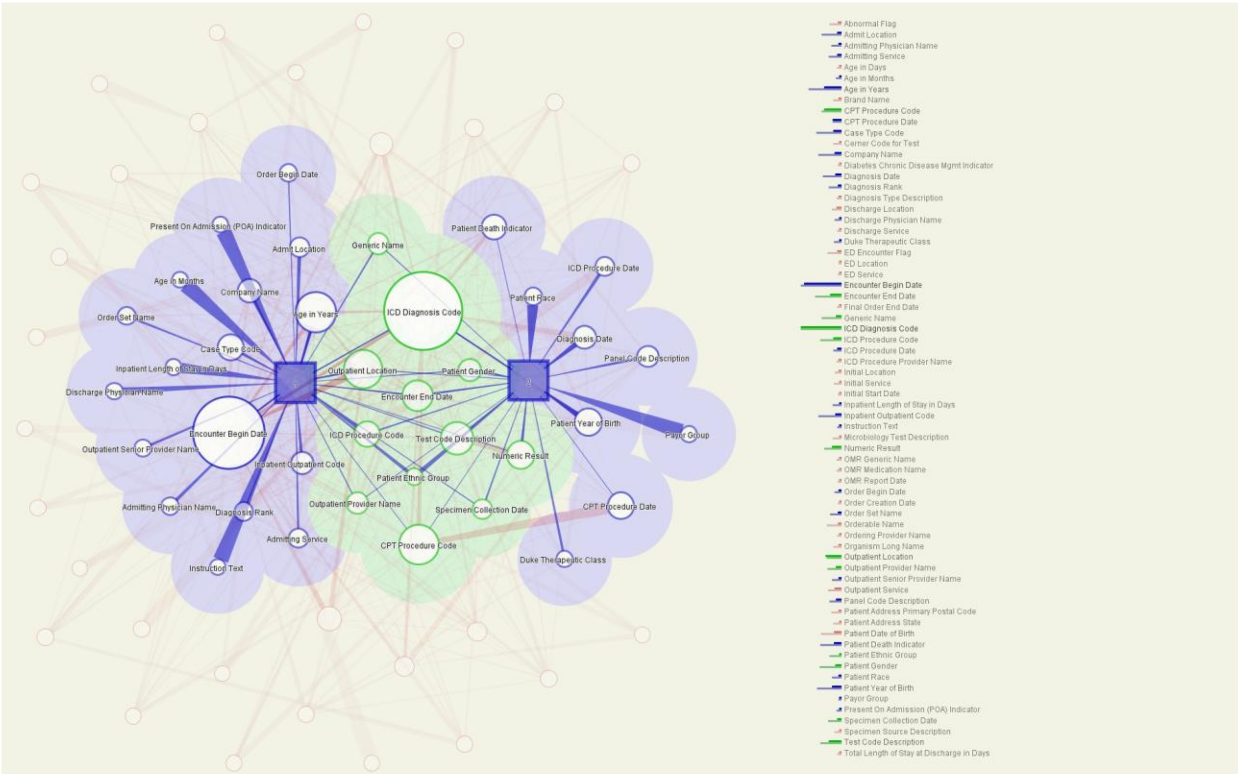


Figure 1. Force-directed network visualization of DEDUCE queries

This visualization shows data elements as circles, and de-identified system users as squares (only the top-two users are shown in this visualization). The node size represents the number of times each node is found across all queries. Links between nodes represent the relative strength of the relationship between the nodes, i.e. how often the nodes were used together in a set of queries. The two users have been highlighted, and circles highlighted in green represent data elements used in queries by both users. Circles highlighted in blue represent data elements used by only one of the users. The desaturated red nodes and links represent data elements not used by either user. This visualization enables the user to interactively explore the connections between users and query terms used in the DEDUCE system.

We have also used our radial coordinates visualization, described in more detail in Section 2.7, to visualize DEDUCE queries. Figure 2 shows the same query data from Figure 1, but with a different representation. 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. With this visualization, we can see which data elements are typically used by each user, and see clusters of data elements that tend to be used together in queries by these users.

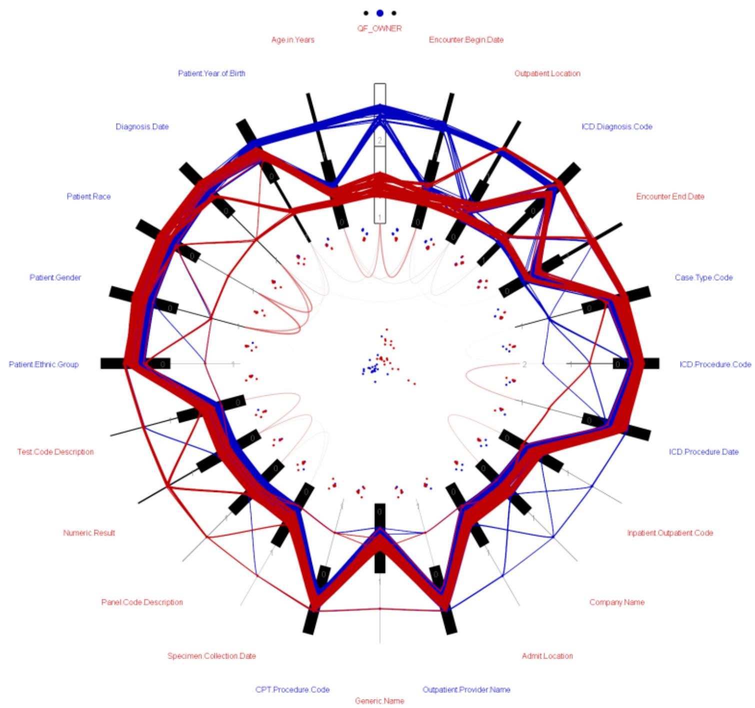


Figure 2.: Radial coordinates visualization of DEDUCE queries.

2.4.1. Explore visualization methods previously applied to health care

Following the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) Statement,[1] a systematic electronic review of the literature was conducted between May and July 2013 to investigate the use of visualization techniques reported between 1996 and 2013. A review using MEDLINE and Web of Knowledge was supplemented with citation searching and a grey literature search. Reference lists from highly relevant articles were also reviewed to find additional articles. Broad key words and search terms were used to assure a comprehensive document search. A matrix was developed for reviewing and categorizing all abstracts and to assist with determining which should be excluded in the review. Articles were excluded if they related to genetics, animals, environment, population health, primarily related to the technical aspects of visualization or position papers, or did not describe specific techniques used for the visualization.

Eighteen articles were included in the qualitative review. (Note: See Appendix A for the schema adapted from the PRISMA group.) Although there is increasing interest in visualization using health care data, in particular population data from EHRs, its use is limited. A manuscript on this review and findings is in process for submission to a Call for Papers for a JAMIA Special Issue of Visual Analytics in Healthcare, due 1 May 2014.

2.4.2. Conduct semi-structured interviews with clinicians to determine how the user intended to use the data from the query, the relevance of the query, and the clinicians' satisfaction and use of the information derived by the query.

Because we were conducting a survey of DEDUCE users, we included questions in the survey that would also give us the information for completing this task. The survey and results are described in Section 2. 2.1.

2.4.3. Develop a Follow-up Questionnaire using a 5 point Likert scale to be used in researcher evaluation of different visualization methods.

One question we also asked in our survey is if the respondent would be willing to provide feedback on various visualizations, and if so, their contact information. We have contact information for 36 respondents who are willing to provide feedback regarding the usefulness of various ways to visualize data when we are ready to begin this task.

2.4.4. Combine interview data with Questionnaire results to evaluate clinical relevance of the visualization methods.

Work has not yet begun on this task.

2.5. Modify or revise classification and data elements of queries based on analysis of the relevance of the visualization methods.

STATUS of milestone Work has not yet begun, pending completion of the classification schema.

2.6. Create a matrix of best visualization techniques.

STATUS: Work is ongoing. We have made significant progress in exploring new visualization techniques, as described in 2.6.1. A matrix of best visualization techniques will be completed when we have clinician input regarding the various techniques used.

2.6.1. Explore ways to mix different types of data in visualization

We have explored mixing different types of data and different representations of data in our force-directed network visualization of query data, our radial coordinates visualization, and in our new co-occurrence visualization.

Query network visualization

In the query network visualization, described in more detail in Section 2.4, we combine the force-directed network visualization with a list view that shows the data element name along with two bars indicating the prevalence and connectivity of the data element. These two visualizations are linked, such that highlighting of data elements in one of the views is reflected in the other views.

Radial coordinates visualization

Due to the delay in obtaining HRPO approval to look at clinical data, we used various datasets packaged with the R statistical programming environment, which we also used for developing visualization prototypes. One of the datasets is the *mtcars* data, which includes performance and design data for various automobiles, e.g. miles per gallon, horsepower, and weight. Although not health-related data, this data set includes continuous data, interval data, and categorical data, thus making it a reasonable starting point for working with heterogeneous data. Using this data we were able to look at combinations of the various data types and how relationships can be easily seen across multiple data dimensions.

Figure 2 shows an example of our radial coordinates visualization prototype developed in R, explained in more detail in Section 2.7, applied to the *mtcars* data. In this visualization, each axis represents a measured quantity of each automobile model (e.g. miles per gallon), and each automobile model is represented by a curved line connecting its value across the various axes. By highlighting lines in different colors, relationships between individuals and groups of entities can be seen across the data axes. For example, we can see there is a strong relationship between displacement (disp) and weight (wt). This visualization also incorporates per-axis distribution visualizations based on data type, clustering of similar axes, arcs showing correlations between axes, and scatter plots of data entities.

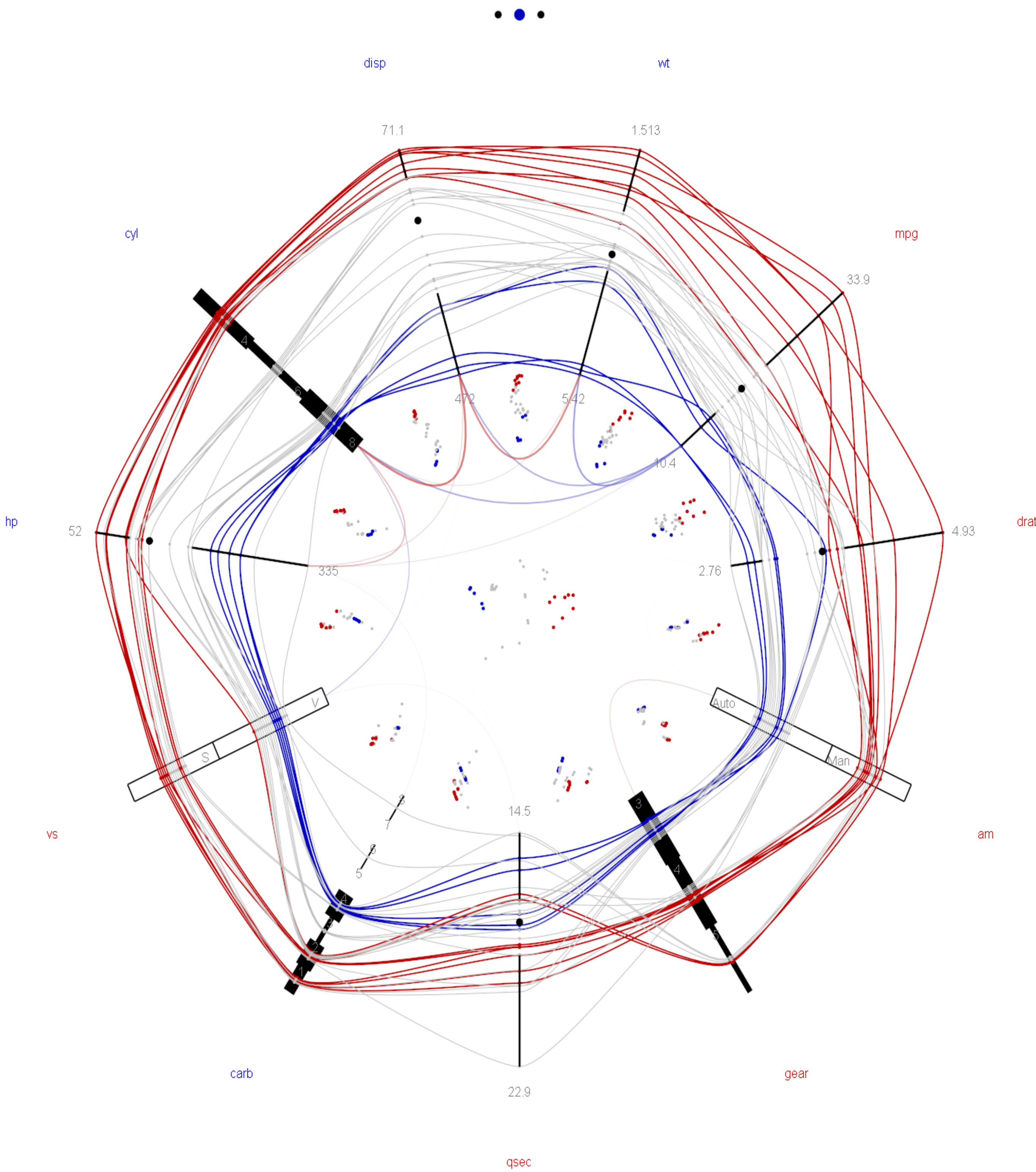


Figure 2. Radial coordinates visualization of mtcars data.

In order to provide improved user interaction, we are currently developing a version of our radial coordinates visualization using the d3 JavaScript library (Figure 3).

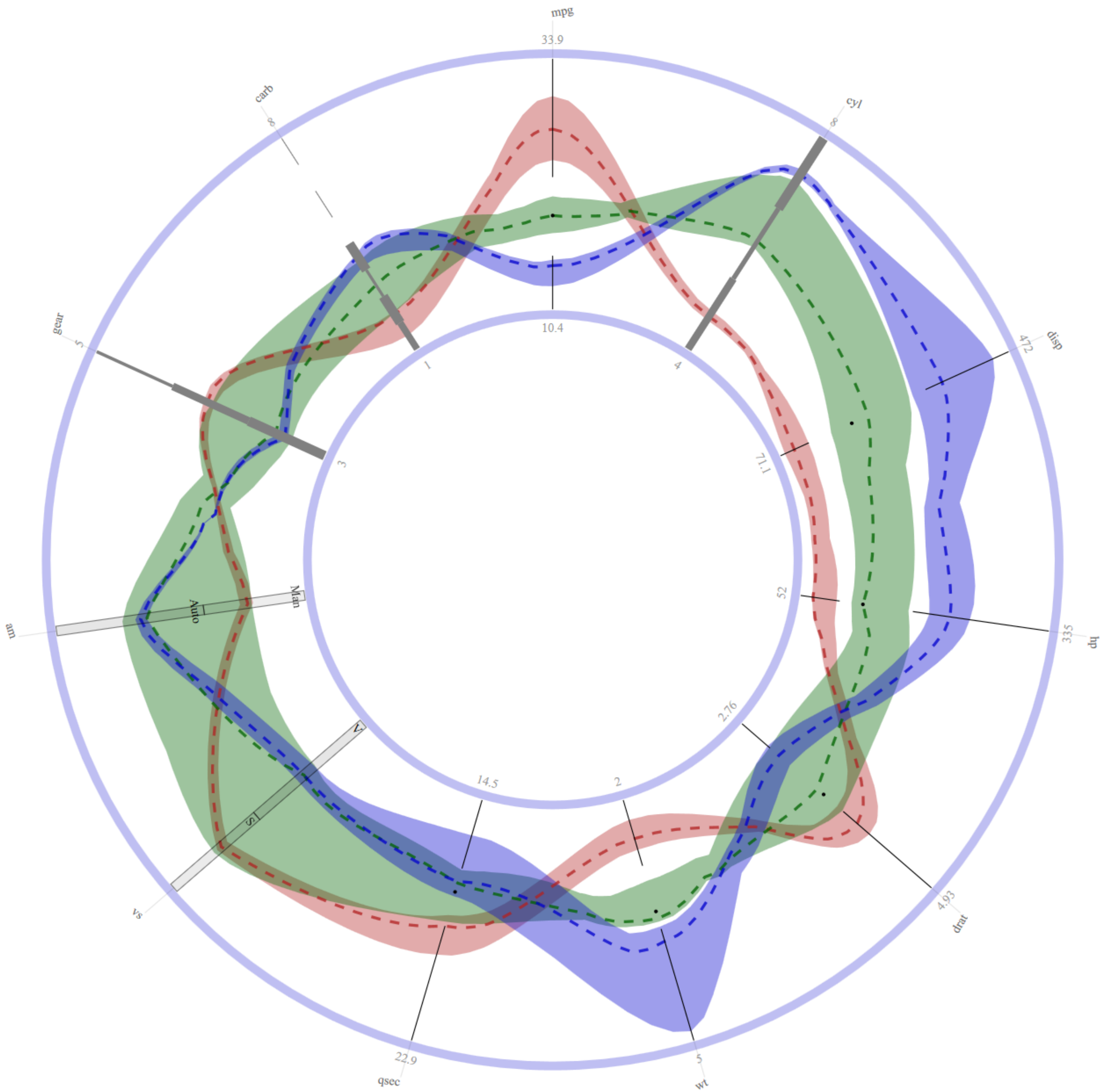


Figure 3: Radial coordinates visualization of mtcars data, showing inner-quartile ranges and median values for three user-selected groups.

One of the key features of our radial coordinates visualization, described in more detail in Section 2.7, is the ability to incorporate multiple types of data, which we currently classify as continuous, discrete, and categorical. Each axis is itself a visualization of the population distribution for that particular variable, with the visualization type determined by the data type. In addition, we incorporate visualization of correlations between axes via clustering and arcs drawn between highly correlated (or anti-correlated) axes. Finally, multiple scatter plots show a different representation of the same data used to generate the lines in the main visualization. Again, selection in one data representation is linked to the other representations.

Co-occurrence visualization

We are currently developing a d3-based co-occurrence visualization tool to investigate the synthetic dataset supplied to us by the DoD. So far we have looked at co-occurrence of problems by patient, medication orders by patient and encounter, lab orders by patient and encounter, vaccines by patient and encounter, and allergens by patient. Figure 4 is an example of this visualization tool applied to problem co-occurrence by patient.

This visualization incorporates multiple linked views: vertical and horizontal bar charts of problem frequency across all patients, a co-occurrence matrix indicating the number of times any two problems co-occur, and a network visualization showing explicit connections between problems over a user-controlled threshold, where node size is determined by the frequency of the problem. The bar charts/co-occurrence matrix can be sorted alphabetically, by frequency, by number of connections, and by the strength of the maximum connection for each problem.

In this example, the user has highlighted Hepatitis B. All other problems that are connected to Hepatitis B are drawn in red, and the problem name is shown. Problems connected to Hepatitis B that are over the user-controlled threshold are drawn more opaquely. Links are drawn between any nodes that are over the user-controlled threshold. In this case, we can see that Hepatitis B is among the most frequent problems in the database, that it is connected to many other problems, and that it shares strong connections to Tuberculosis, Joint Paint, and Shortness of Breath. Because we are working with synthetic data, we are cautious in making new knowledge decisions, which is a temptation in using visualization to understand big data.

Future work on this tool will investigate the ability to incorporate temporal relationships between data elements, look at the relative strength of co-occurrences modulated by data element frequencies, incorporate hierarchical structures of data elements (e.g. grouping problems based on ICD codes), and look at comparisons between user-selected groups of data elements.

Figure 4: Co-occurrence visualization of patient problems, with problem frequency bar charts, co-occurrence matrix, and network visualization. Hepatitis B occurs quite frequently, and co-occurs with many other problems.

2.7. Develop parallel coordinates visualization of data resulting from data queries.

STATUS of milestone: Work is ongoing.

We have developed two radial coordinates visualization tools based on parallel-coordinates [2,3] and star plot[4] multivariate visualization techniques. The first prototype was developed using the R statistical programming environment, and the current version is created using the d3 JavaScript visualization library.

R-based radial coordinates

R is a popular programming environment for statistical analysis, including the generation of statistical graphics. We therefore investigated the use of R for developing visualization prototypes, enabling our research team to work together more closely and generate prototypes more rapidly. Additionally, R includes numerous data sets that were useful for prototyping purposes before we had HRPO approval to look at clinical data. To enable some degree of interactivity, we used the rgl library, which provides an interface to hardware-accelerated OpenGL graphics within R.

In our radial coordinates prototype, each data entity (e.g. an individual patient) is represented by lines connecting that data entity's measured value for each axis (e.g. height, weight, race, age, sex, etc.). Our radial coordinates visualization prototype incorporates several enhancements, including:

- Radial axis layout providing a square aspect ratio, which can be beneficial for large numbers of axes;
- Axis distribution visualizations based on numeric type (continuous, discrete, categorical);
- Line spreading for integer and categorical data[5] , mitigating the problems of multiple lines collapsing to a single data point for discrete and categorical data, extending the parallel sets method[6] to enable visualization of individual data entities, and the incorporation of non-categorical data;
- Curved lines to make it easier to visually track along lines;
- Automatic axis clustering based on correlations between axes;
- Direct visualization of axis correlations via colored arcs connecting axes;
- Automatic optional axis flipping based on correlations to minimize line crossings;
- Incorporation of pairwise scatterplots for neighboring axes and a central scatterplot based on the first two principal components;
- Interactive line brushing to highlight groups in different colors;
- Interactive coloring by axis.

Figure 5 shows an example of the radial coordinates technique applied to Primary Care Trust (PCT) data from the United Kingdom's National Health Service (NHS). 147 of the 152 PCTs in

England are represented (only PCTs where data was available for all axes are included). In this visualization, the user has selected the *lung_Combined_DSR* axis (circled), a measure of the prevalence of lung cancer in each PCT. In this manner, the user can see how the PCTs with high and low lung cancer rates are distributed across the other axes. For example, we can see some interesting clustering of the red and blue lines along the *ONS_Area_Class_Group* axis (circled). The ONS is the Office for National Statistics, and this axis represents a classification of the region serviced by the PCT (e.g. London Suburb, Prospering Smaller Town, Industrial Hinterlands, etc.).

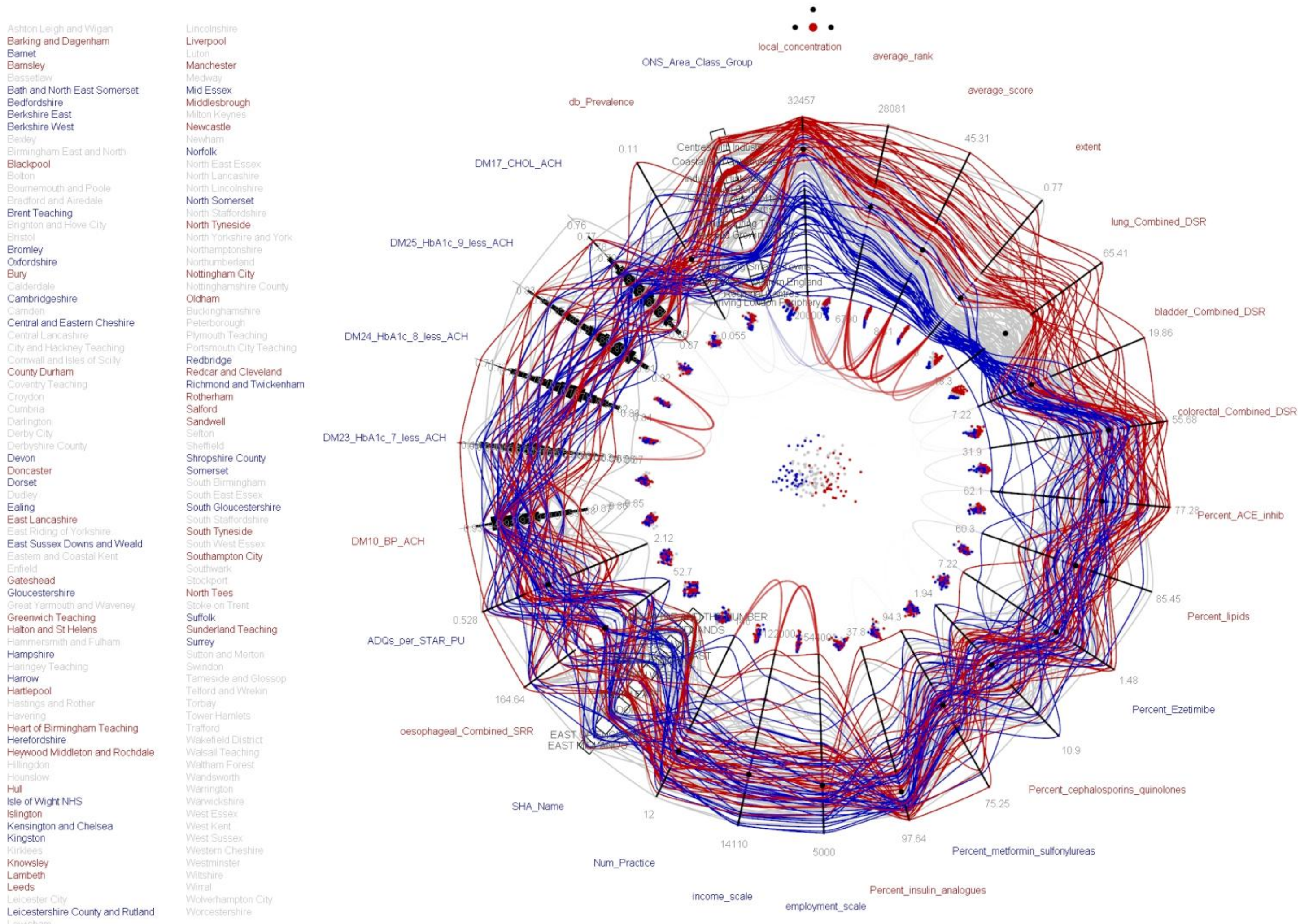


Figure 5. Radial coordinates visualization of NHS Primary Care Trust (PCT) level data, colored by lung cancer prevalence (red: high, blue: low).

By zooming in (Figure 6), we can see that the red lines (high lung cancer) tend to cluster in Industrial Hinterlands, Centres with Industry, Regional Centres, and Manufacturing Towns, whereas the blue lines (low lung cancer) tend to cluster in Thriving London Periphery, Prospering Southern England, and Prospering Smaller Towns.

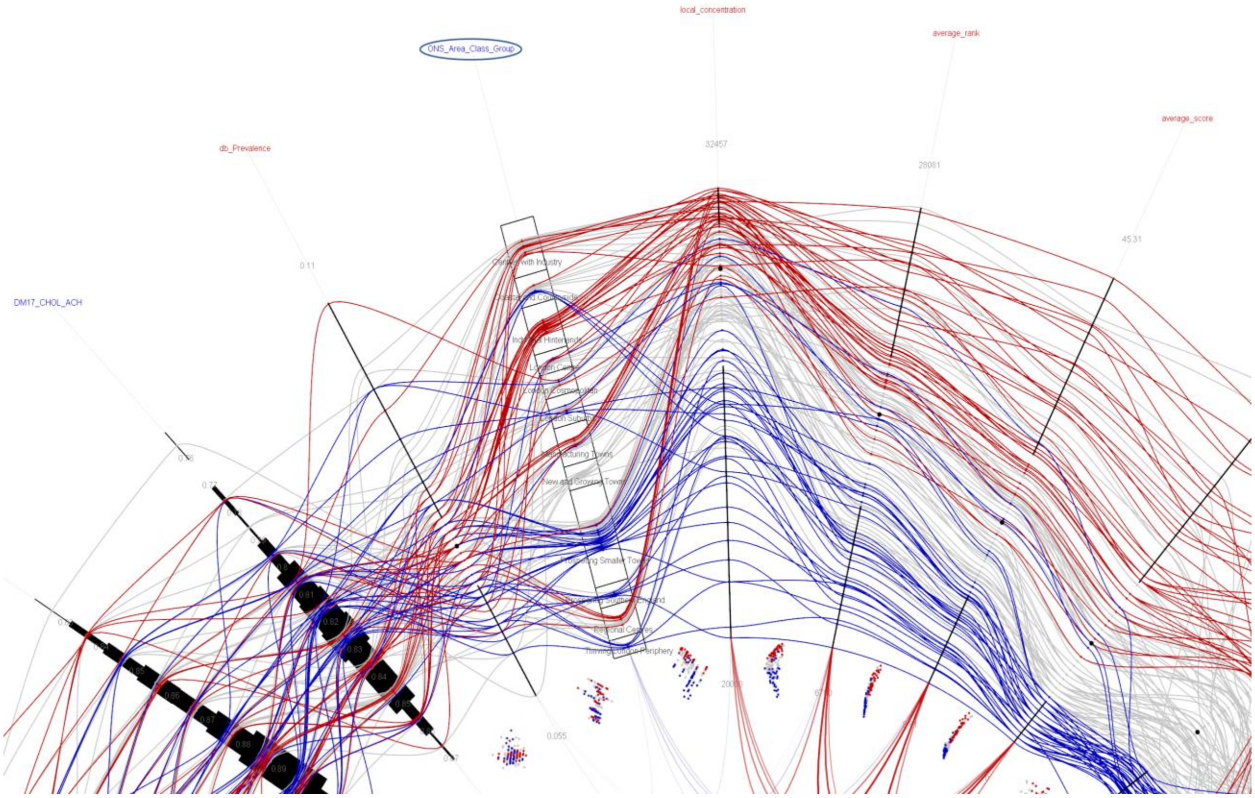


Figure 6. Radial coordinates visualization of NHS Primary Care Trust (PCT) data, colored by lung cancer prevalence (red: high, blue: low). Clusters of high (e.g. Industrial Hinterlands), and low (e.g. Prospering Smaller Towns) lung cancer rates are visible for various categories of the "ONS_Area_Class_Group" axis.

Although the R-based radial coordinates prototype proved very useful, especially with respect to the ability to incorporate statistical methods such as correlations and principle components analysis, some drawbacks were determined. These include the lack of support for blended transparency (a common technique for dealing with over plotting in parallel coordinates techniques), and difficulty in incorporating more advanced interactions.

d3-based radial coordinates

Although our initial plan was to generate prototype visualizations using R and Processing, before creating applications based on the Visualization ToolKit (VTK), we have decided instead to use the d3 JavScript library for future development work. This decision is based in part on recommendations from colleagues at the AMIA 2013 VAHC workshop. d3 is quickly becoming the de facto standard for web-based visualization, and is designed to enable high interactivity. Additional benefits include the ability to incorporate other web-based solutions for interfacing with large data sets. One potential drawback of d3 is the lack of statistical methods present in the R programming environment. However, by utilizing the Shiny package, which enables interactive web-based R applications, we plan on combining the powerful visualization capabilities of d3 with the statistical capabilities of R.

Our d3-based radial coordinates visualization tool does not yet include some features of the R-based prototype, including correlation-based axis clustering, direct visualization of correlations via curved arcs, and scatterplots, in large part because we do not have direct access to the statistical calculations available in R, such as correlations and principle components analysis. These features will be added, either via exporting such data from R and loading separately, or via a direct connection to R using the Shiny library. Thus far we have concentrated on improving the user interface to the radial coordinates visualization and improving the ability to convey summary statistics of selected groups. These improvements include:

- User control of various parameters, such as opacities, discrete value threshold, and curve properties;
- Improved selection capabilities. Users can now easily add to selections, subtract from selections, and select by clicking on various parts of the visualization, such as labels and axis parts; and
- Per-group "ribbon" and summary statistic overlays (described in more detail in Section 2.7.1).

Figure 7 gives an example of our d3-based radial coordinates visualization applied to the NHS PCT data. The user has selected a specific London suburb, Harrow, highlighted in red, and has then selected all other London suburbs, highlighted in blue, by changing the selection color and clicking on the *London Suburb* label on the *ONS_Area_Class_Group* axis. In the image on the left, it is possible to compare Harrow to the other London suburbs, but the visualization on the right makes this easier by reducing the opacity of unselected PCTs, and drawing a ribbon outlining the maximum and minimum values for the blue group. The user can then easily see where Harrow is similar to the other London suburbs on aggregate, and where it differs.

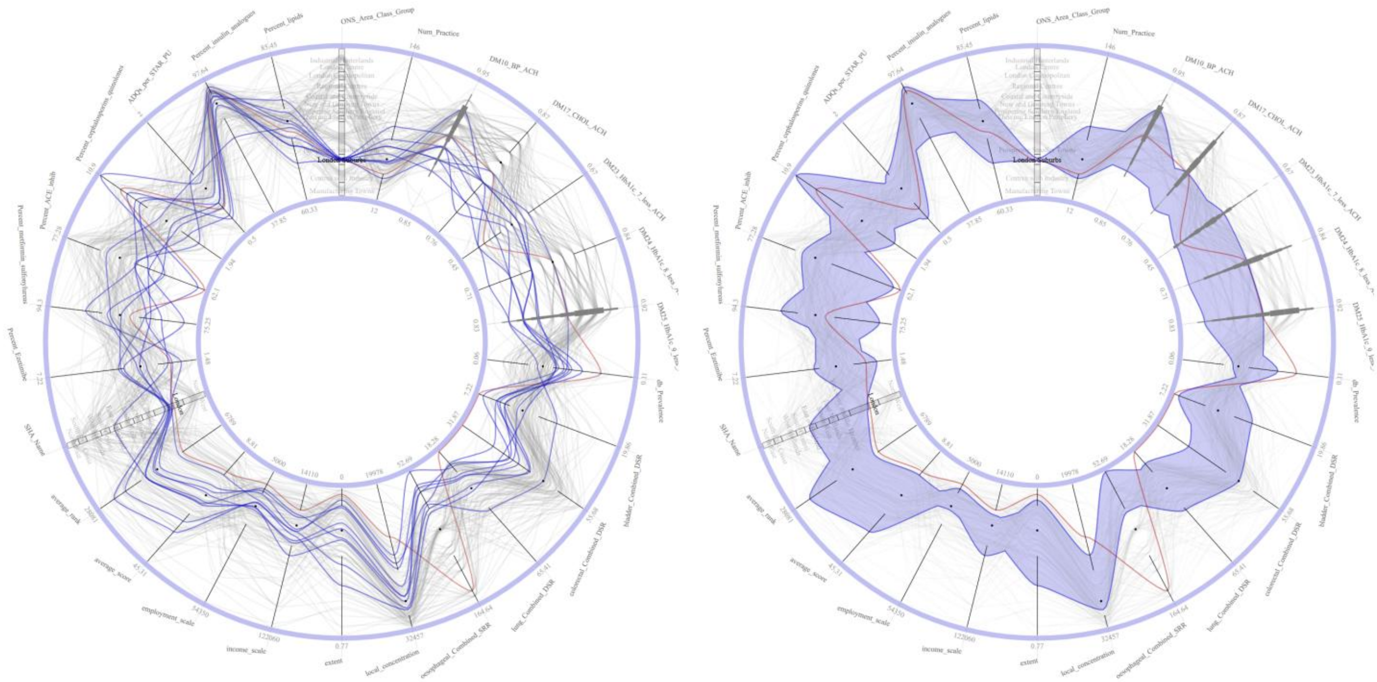


Figure 7. Radial coordinates visualization of NHS Primary.

The largest distinction can be seen in the lower right of the visualization (Figure 8). Harrow has the lowest across the entire data set for three consecutive axes related to lung, colorectal, and bladder cancer rates. However, Harrow has a much higher esophageal cancer rate than the other London suburbs. This ribbon visualization technique is useful for comparing the spread of different user selected groups versus the overall population distribution represented by each axis.

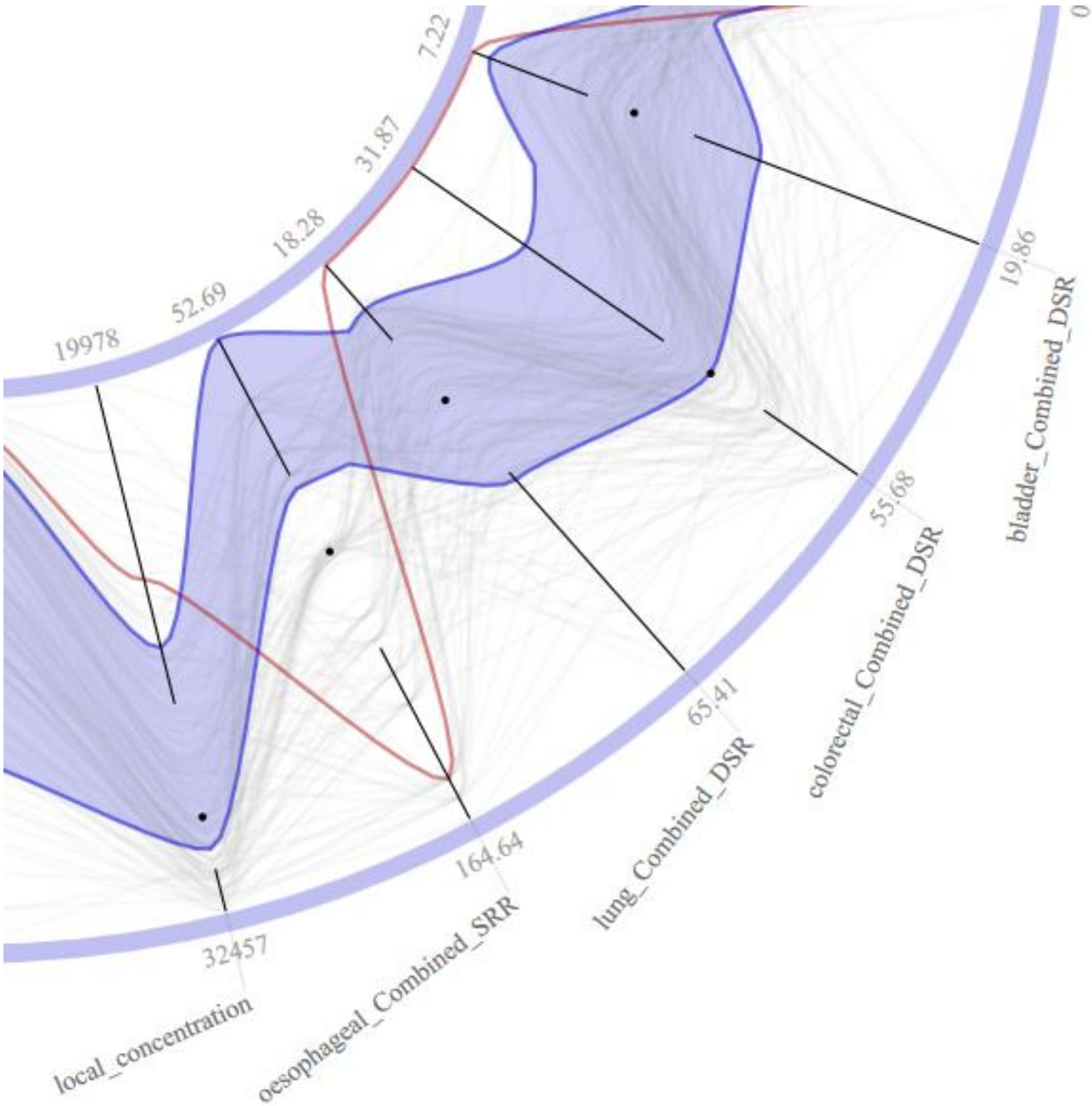


Figure 8: Close up of radial coordinates visualization, highlighted the discrepancy between the Harrow PCT (red) and the other London suburbs (blue) with respect to the *oesophageal_Combined_SRR* axis.

2.7.1 Extend parallel-coordinates visualization to include summary statistics per data element and evaluate its ability to reveal significant patterns.

Our radial coordinates visualizations incorporate axis distribution visualizations based on numeric type (continuous, discrete, categorical). Figure 9 gives an example of each type of axis distribution visualization.

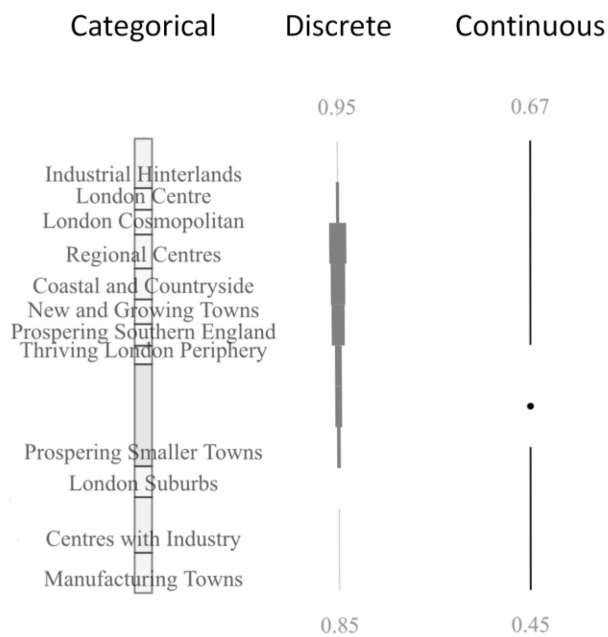


Figure 9: Categorical, discrete, and continuous axis distribution visualizations.

Categorical axes are represented by a stacked bar chart, with the height of each section proportional to the number of data elements with that value. Discrete axes are represented by a histogram with bin width proportional to number of entities with that value. Continuous axes are represented by a box-and-whiskers plot showing the four quartiles of the data distribution.
Each axis is therefore a visualization of the population distribution for that variable.

To enable improved comparison of the distribution of user-selected groups, we have enhanced our ribbon visualization technique with summary statistic overlays. Figure 10 on the next page shows an example of this technique, comparing industrial hinterlands (red) with prospering southern England (blue) from the NHS PCT data. In this visualization, individual PCT lines are not shown. Instead the faint red and blue section show the overall range of the data values for each group, the dark red and blue sections show the inner quartile ranges for each group, and the dashed line shows the median for each group. In this manner we can directly compare the data distributions between each group, and against the overall population.

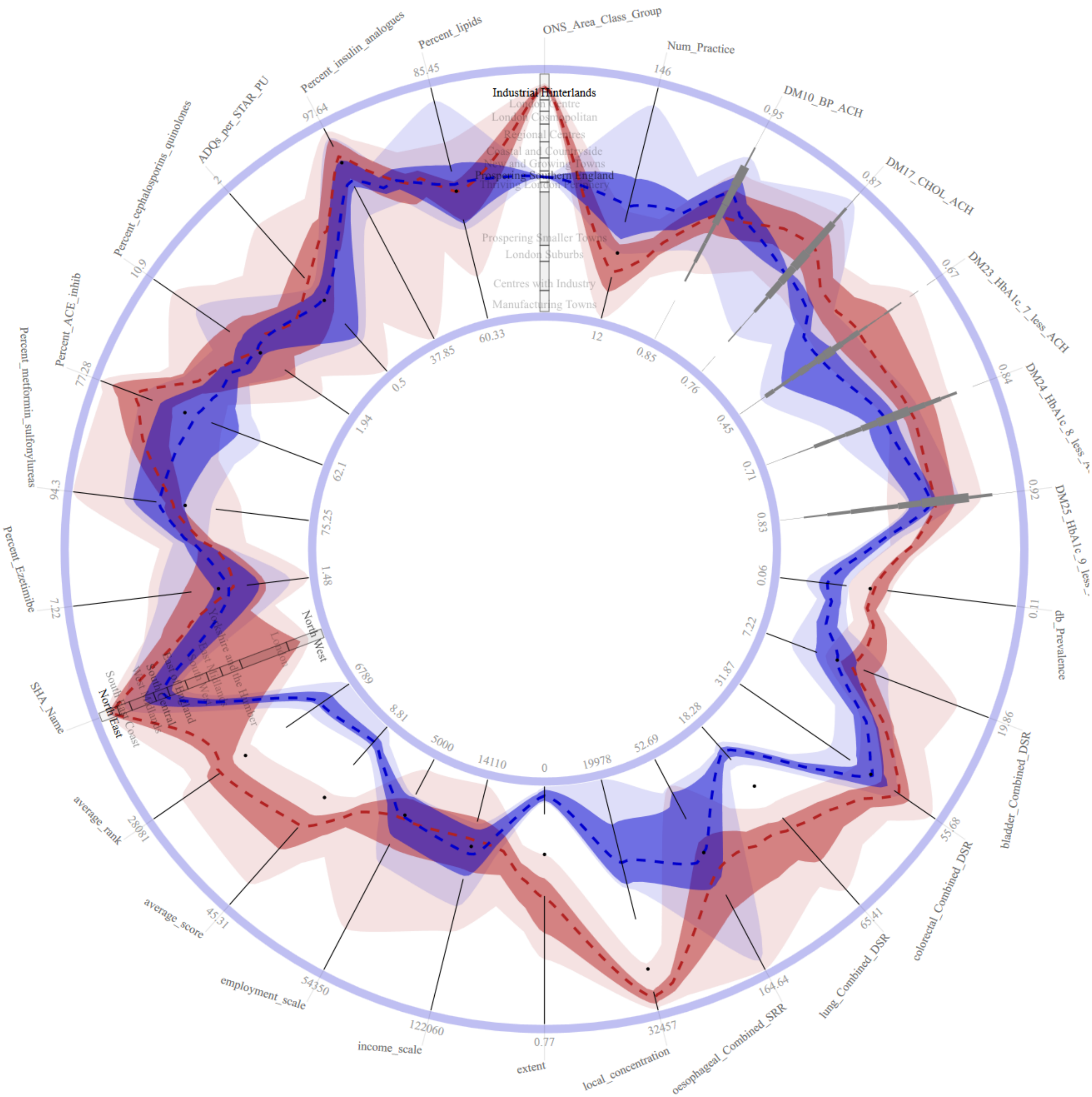


Figure 10. Ribbon visualization technique with summary statistic overlays comparing *Industrial Hinterlands* (red) with *Prospering Southern England* (blue) from the NHS PCT data.

2.8. Add supplemental data-dependent linked views of the data based on matrix of visualization techniques.

STATUS of Milestone: Work is in process.

We have incorporated linked views into all visualization tools developed so far. The DEDUCE query visualization a force-directed network visualization for its main view, with a supplemental list/bar chart view indicating frequency and connectivity for each node. Nodes can be selected in either view, and these selections are reflected in both views.

In addition to the main radial coordinates view in the R-based prototype, multiple scatter plots are shown. Selection in either the radial coordinates or any scatter plot is reflected in the other views.

The co-occurrence visualization incorporates bar charts, a co-occurrence matrix, and a force-directed network visualization. User selections in all views are reflected in the other views.

2.9. Complete testing visualization using PTSD.

STATUS of milestone: Work will begin towards the end of the project.

2.10. Future Work.

We are now focused on identifying Phenotypes for diseases. We propose to use data visualization to identify phenotype signatures for a disease. Using the technique already discussed, we plan to look at the frequency of occurrence of data elements associated with a specific disease that are not associated with individuals not having the disease. We then plan to do the opposite – looking across the aggregated databases identifying the frequency of occurrence of related data elements for those not having the disease that do not appear for those having the disease. From this dual approach, we will create a computable algorithm, identifying the data elements that are present and those data elements that are not present. Using frequency of occurrence as a weighting facture, we propose that we can create weighting factors that permit us to calculate a certainty factor for the presence of a disease in an individual based on the presence or absence of date element. For quantitative data elements, we will subset the continuous range of values into groups that provide the maximum separation.

One value of this approach is that if the data elements are present that result in a certainty factor of 100%, then the source of the data will become less important. This approach will create a measure of quality as well as establishing trust in data coming from other sites.

During our initial work we used the R statistical programming environment to generate prototype visualizations, which although useful, suffered from difficulties in incorporating sophisticated user interactions. We have therefore begun using the d3 JavaScript library, which is designed to enable dynamic visualizations based on user interactions. One drawback of d3 is that it does not include many of the statistical capabilities of R, however going forward we plan to incorporate the Shiny R library, which enables web-based interaction with an R server. Combining d3 and R in this manner should enable a powerful visual analytics framework.

3. KEY RESEARCH ACCOMPLISHMENTS

The key research accomplishments emanating from this research to date are as follows.

- Systematic review of the literature to identify how visualization is used with health care data.
- Identification of the data elements most frequently requested by users with access to "big data" from electronic health records.
- Development of R-based radial coordinates multivariate visualization prototype.
- Development of Processing-based force-directed network visualization of DEDUCE query data.
- Development of d3-based radial coordinates visualization with improved interactivity.
- Development of d3-based co-occurrence visualization tool, applied to synthetic data obtained from the DoD.

4. REPORTABLE OUTCOMES

Results from this research to date include the following reportable outcomes.

4.1. Presentation

DEDUCE was developed as an online research tool in 2008 for use by Duke investigators who conduct queries on the clinical information collected through patient care activities. A guided query is used to filter through the millions of rows of data to obtain information researchers seek for an Internal Review Board (IRB) approved research activity. A DEDUCE Users Group meets at Duke monthly for updates on the tool, and to discuss problems and solutions. We gave a demonstration of the visualization techniques and results from our earliest visualization, which used counts of the various data elements queried using DEDUCE.

4.2. Abstract and podium presentation

In its fourth year, the 2013 Workshop on Visual Analytics in Healthcare (VAHC 2013) was held in conjunction with the American Medical Informatics Association (AMIA) Annual Symposium in Washington, DC 16-20 Nov 2013. This day long Workshop provided an opportunity for participants to discuss visualization techniques, software applications, and datasets that are being used in various health care settings. We submitted two abstracts for peer review. Our first abstract, entitled "Visualization of EHR and Health Related Data for Information Discovery," was accepted as a podium presentation. Please see Appendix B for the abstract.

4.2 Abstract and demonstration

During VAHC 2013, a second peer-review abstract was accepted as a demonstration for the afternoon session. During the demonstration, we showed interested participants the data visualization techniques we have used on two of the data sets we worked with prior to HRPO approval: the query data elements data and data from British Telecom. Please see Appedix C for the abstract.

5. CONCLUSION

Although tasks as planned on this project were to begin using clinical data from DEDUCE, we were delayed and did not progress according to our timeline. We were able to find and use other data to test visualizations, however, and conducted a survey to evaluate the type of data users seek with access to large amounts of electronic data and why queries are conducted. These steps provided us with information regarding visualization of data that has been very useful as we began working with actual clinical data the last quarter of 2013. The objective of our project is to explore interactive visualization of large sets of health data to provide better understanding of what is in the data. An interface that allows a user to interactively explore various data elements, using petabytes of health care data representing many data elements compressed into various groups of related data, and presenting this visually to the user, has the potential to be an important means to gathering information about large amounts of data in electronic records.

Our hypothesis is that data visualization is more effective than traditional methods of data exploration, and that this type of visualization is highly dependent on the data and nature of the queries and what someone is trying to learn. From our work using the NHS data, we were able to discern important information that, through examination of the literature, has previously been reported. The incidence of esophageal cancer is highest in the Harrow PCT, which is a population primarily of people from India; Indians are reported to have high rates of this type of cancer. There are numerous ways this information might be used in population health strategies or practitioner sites. The potential to detect causal relationships between various sets of data may lead to improved health care and resiliency in not only military personnel but all whose data are included in the visualization of aggregated data. It could also assist the Department of Defense in strategic decisions about personnel, perhaps identifying certain people who might, for instance, be poor candidates for certain geographical locations (e.g. areas with high pollen that increases the need for more intensive health care for people with asthma) and save millions of dollars of health care costs by early identification of vulnerable populations.

In the future, we propose to quantitatively measure the value of adding certain other data into the EHR for use with our approach. For example, does environmental data increase the certainty of certain diagnoses? Does aggregation of patient data across sites of care increase the certainty value of diagnoses? What do genomic data and biomarkers add to the diagnosing certainty? What do patient reported outcomes contribute to diagnosing, determining the correct treatment, and caring for a patient? We propose that future research is needed to understand how to assign weighting factors and how to use this approach in a patient centric environment

REFERENCES

1. Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, 151(4), 264-269.

2. d'Ocagne M. Coordonnees parallels et axiale. Gautier-Villars, Paris 1885.

3. IInselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer, 1*(2), 69-91.

4. Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. Graphical methods for data analysis. 1983. Wadsworth, Belmont, CA.

5. Graham, M., & Kennedy, J. (2003, July). Using curves to enhance parallel coordinate visualisations. In *Information Visualization, 2003. IV 2003. Proceedings. Seventh International Conference on* (pp. 10-16). IEEE.

6. Bendix, F., Kosara, R., & Hauser, H. (2005, October). Parallel sets: Visual analysis of categorical data. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on* (pp. 133-140). IEEE.
.

APPENDICES

Appendix A: Flow of information during a systematic literature review

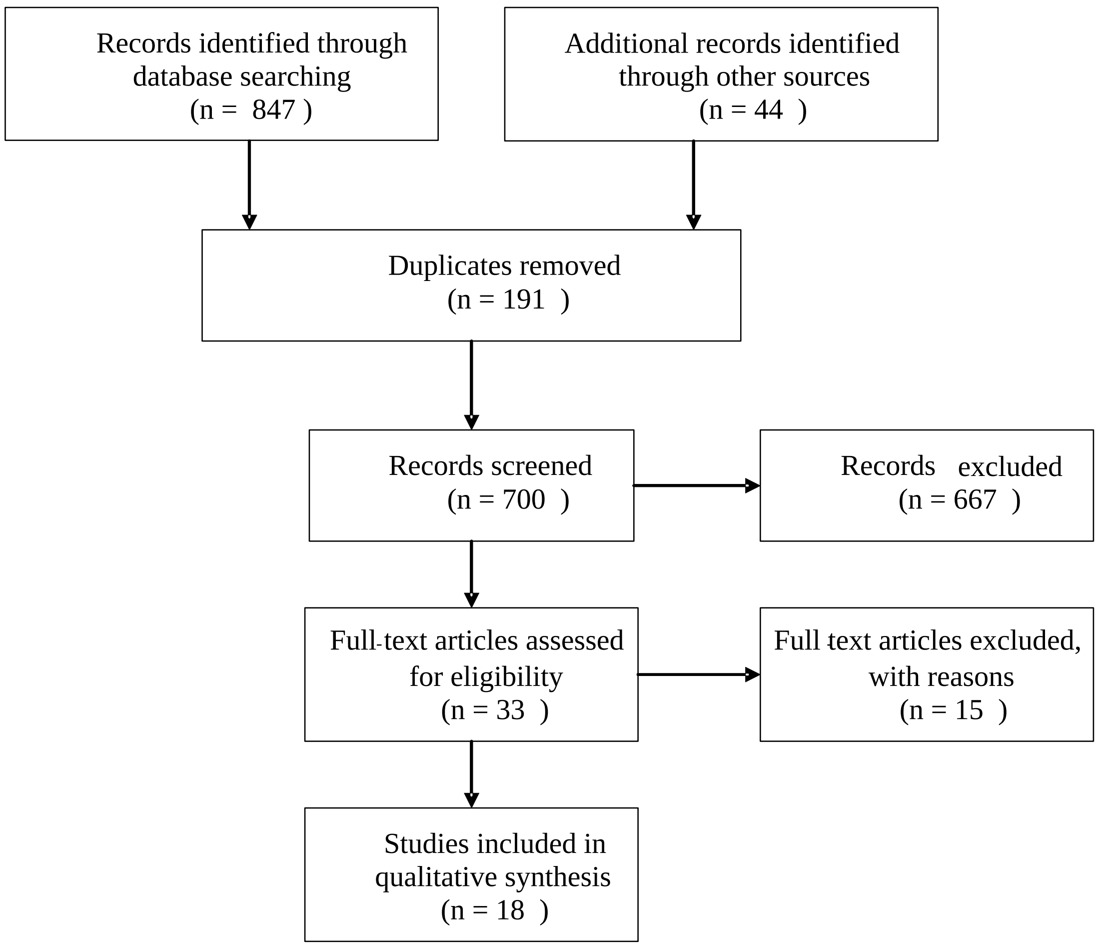
Appendix B: Survey on the Use of DEDUCE queries

Appendix C: Reprints of peer-reviewed abstract as a podium presentation

Appendix D: Reprint of peer-reviewed abstract presented as a demonstration

APPENDIX A

Figure 1. Flow of Information through the Different Phases of Systematic Review. Adapted from The PRISMA Group, Moher D[1].



APPENDIX B

Survey on the Use of DEDUCE Queries

We are conducting an in-depth review of data queries to identify what data elements are included in queries, which will be used as a means to explore novel visualizations of large health data sets. We expect this approach to digitized healthcare data will lead to effective visualization of data, with an understanding of aggregated data that leads to discoveries within the data that would otherwise not be possible.

To help with this exploration, we hope you will be willing to share your thoughts and ideas with us about your use of the DEDUCE query system. Your responses will be aggregated with other respondents, and your name will not be used in any of the reports that might arise from this survey. Survey responses will be used to help us understand what information clinicians seek from data available to them, which will hopefully help us find the most effect ways to visualize the information.

1. Do you use DEDUCE for data queries?
   - Yes
   - No

2. Do you run your own queries?
   - Yes, always
   - Yes, most of the time
   - Sometimes
   - Seldom
   - No, never

3. Do you have someone else run the queries for you? If so, who?
   - No / Yes, followed by:
   - Clinical coordinator
   - Fellow
   - Clinical manager
   - Nurse Practitioner
   - Physician Assistant
   - Someone was trained to run all department's queries

4. Approximately how many times have you initiated a DEDUCE query in the past 2 years?
   - >20
   - 15-19
   - 10-14
   - 5-9
   - 1-4
   - 0
   -

5. Why do you run queries? Check all that apply.
   - Thinking about writing a grant but need to know prevalence in Duke patient population
   - Need information for writing a grant
   - Need to see if there are enough patients who will meet inclusion criteria to participate in an industry-sponsored clinical trial
   - Searching for treatment methods
   - Searching for outcomes
   - For quality improvement
   - Other clinical reasons: ____________
   - Additional reasons other than clinical reasons:_______________

6. What information do most of your queries seek? Select all that apply:
   - Demographics
   - Vital signs
   - Diagnoses
   - Medications
   - Procedures
   - Laboratory data
   - Imaging data
   - Device information
   - Geospatial information
   - Encounters
   - Physicians
   - Text for analysis
   - Other. Please list anything not on this list.

7. Do most of your queries provide you with the kinds of information you were looking for?
   - Yes, always
   - Yes, sometimes
   - No

8. Was the information you were seeking available with the first query that was run? If not, approximately how many times do you revise most of your queries before you get the information you want?
   - Almost always satisfied with the first query done
   - Every query usually needs to be revised
   - Usually revise once
   - Usually revise twice
   - Usually revise 3-4 times
   - Usually revise 5 or more times

9. In what format was the query information first presented to you?
   - Excel table
   - ASCII file

- Bar graph
- Line graph
- XY graph
- Other. Please list

10. Did you change the information to another format? If so, what did you use?
    - Excel table
    - ASCII file
    - Bar graph
    - Line graph
    - XY graph
    - Other. Please list all you have used.

11. Would you be willing to provide feedback to us in the future regarding the usefulness of various ways to visualize data?
    - No
    - Yes
    - If yes, contact information:
        - Name
        - Email address
        - Phone number

Appendix C

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.

Appendix C

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.

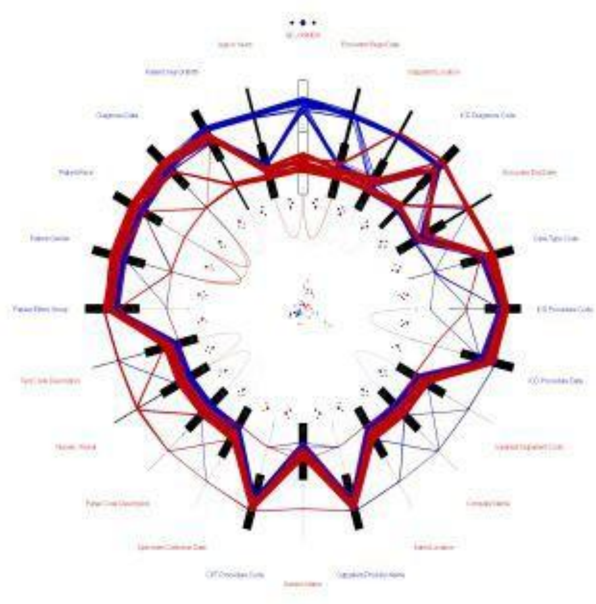


Figure 1: Radial coordinates visualization of DEDUCE queries.

Appendix C

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.

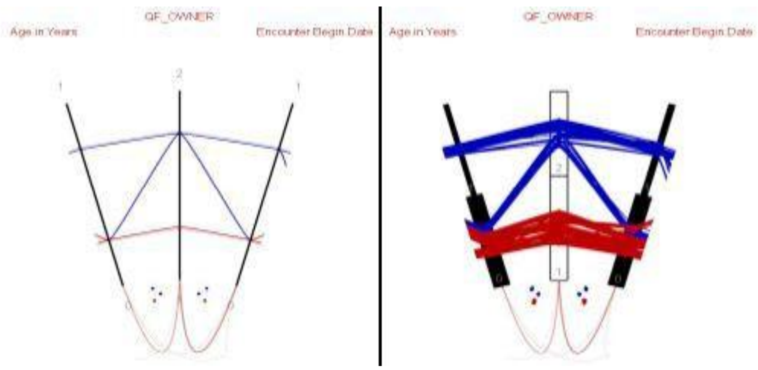


Figure 2: Line spreading (right) for discrete and categorical data enables improved visualization of multiple entities with the same value.

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

Appendix C

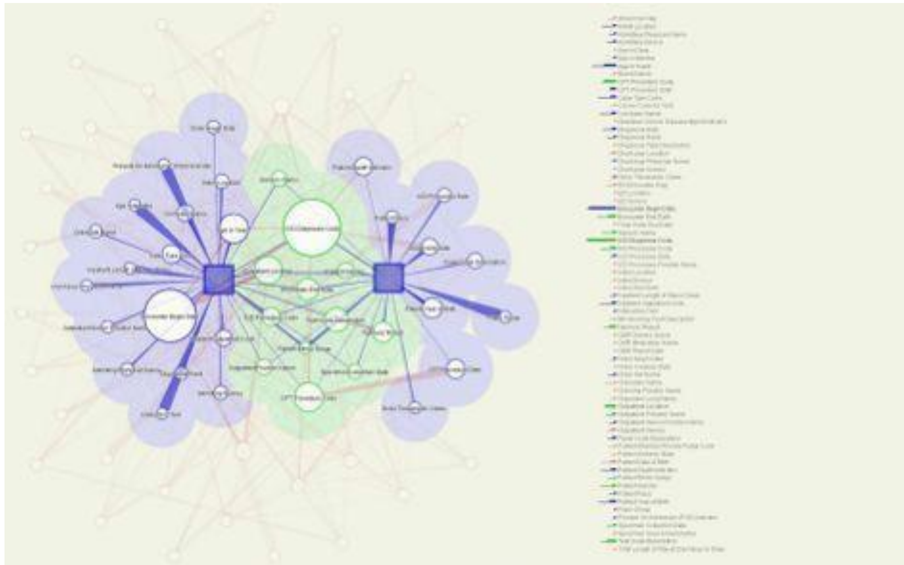


Figure 3: Force-directed layout visualization of DEDUCE queries.

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.

Appendix C

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.

APPENDIX D

Demonstration #3

Demonstration of Visualization of EHR and Health Related Data for Information Discovery

David Borland[1] Vivian West[2] W. Ed Hammond[2]

[1] Renaissance Computing Institute, The University of North Carolina at Chapel Hill

[2]Duke Center for Health Informatics, Duke University

Introduction

In this demonstration we present research we are conducting in response to a program announcement solicited by the Assistant Secretary of Defense for Health Affairs, Defense Health Program. We have developed visualization prototypes for multivariate heterogeneous data along with visualizations of retrospective data queries from DEDUCE, an electronic health record (EHR) query tool developed at Duke University.

Our current approach involves incorporating data queries of Duke's EHR system to help identify what data elements are used in queries and classify them according to what types of information users were seeking (e.g. queries searching for outcomes, or outliers of treatment methods). Eventually 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.

Methods

We have developed two interactive visualization prototypes, one a radial coordinates visualization (Figure 1, left) based on parallel coordinates techniques, and one a force-directed node-and-link network visualization (Figure 1, right). Our radial coordinates visualization is a multivariate visualization suitable for heterogeneous data that incorporates multiple supplemental scatterplots, direct visualization of axis correlations, and a novel technique for spreading lines to enable improved visualization of individual lines and line clusters. Our force-directed network visualization enables the interactive selection of nodes to see relationships between groups of nodes.

In our demonstration we will show how various relationships in the data are reinforced between the two views, and how different visualizations can be more adept at showing different relationships in the data. We will also apply these visualization techniques to publicly available EHR data from the NHS.

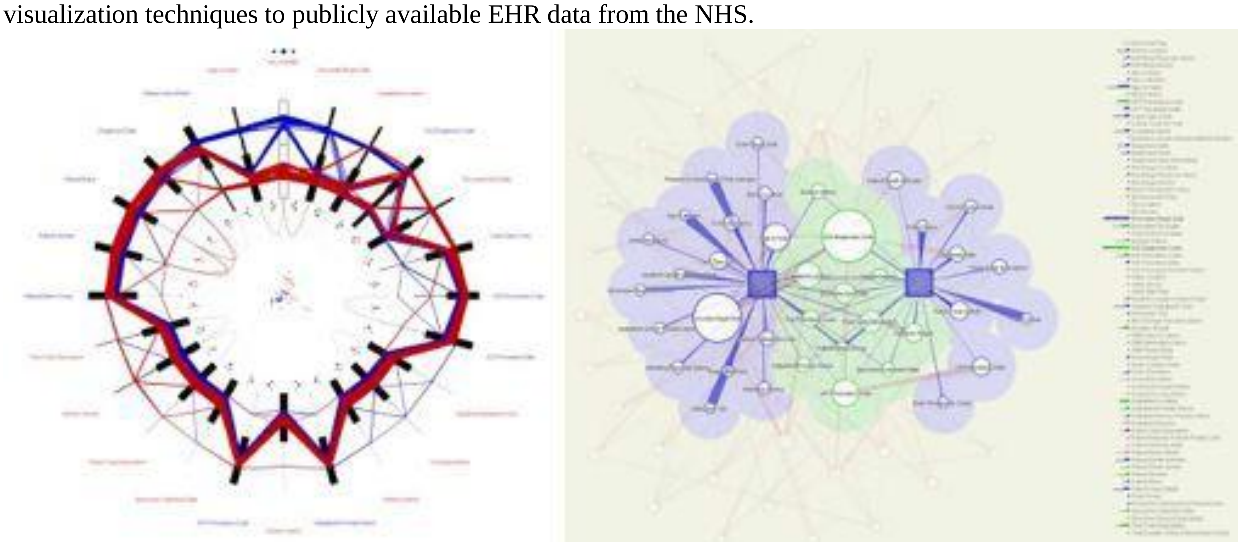


Figure 1: Radial coordinates (left) and force-directed network (right) visualizations of the same EHR query data

Acknowledgements

This work is supported by the US Army Medical Research and Materiel Command (USAMRMC) under Grant No. W81XWH-13-1-0061. The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.