HealthCorr: A Geospatial and Multivariate Data Visualization Tool for Healthcare Datasets

Paul Rosenfeld*, Ginnah Lee*, Awalin Sopan**, Angela Song-Ie Noh**, Sohit Karol***
Department of Electrical and Computer Engineering*, Computer Science**, Kinesiology***
University of Maryland, College Park
{prosenf1, ginnah, awalin, angela, skarol} at umd dot edu

Abstract—Millions of dollars are spent each year on Medicare costs and the Department of Health and Human Services keeps track of these spending and their corresponding effects across the country, resulting in a large geospatially multivariate data set. Current visualization tools for Medicare data sets lack the ability to do multivariate comparisons and show the geographic distribution of the selected variables at the same time. Moreover, such tools are not well suited to exploratory tasks and lack the ability to deliver deep insights into this huge repository of data. HealthCorr is a web application designed to visualize Medicare and other health care data in multivariate space as well as geospatially. The tool enables users to visualize the geospatial distribution of a given variable on an interactive map, and compare two or more variables using charts and tables. By employing four dynamic filters, all the visualizations can be narrowed down to specific regions having demographic characteristics of interest. Healthcorr provides a powerful interface policy makers to visualize the current state of Medicare spending, quality of health, and access to care in an effort to help them make informed decisions about how to improve health care in their region.

Index Terms—Health care, Information Visualization

I. INTRODUCTION

As the health care debate rages on in congress, there appear to be numerous opinions about what is wrong with health care in the US. While everyone agrees that the cost of health care is exploding for both individuals and businesses, the reason why costs rise but quality doesn’t necessarily follow suit appears to be a much more open question. This is further complicated by the fact that costs of health care rising much quicker in some parts of the country than others. Portions of the health care system are plagued by waste, inefficiency, and in some cases, fraud [1,2,3]. Medicare is a social insurance program administered by the Health and Human Services (HHS) department of the United States government, and forms a major part of the fiscal expenditure every year. The total Medicare spending reached $440 billion for the fiscal year 2007, or 16% of all federal spending and is projected to rise sharply with the aging population [4]. With this amount of money at stake, it is imperative that the money being spent on such initiatives is used efficiently and effectively. Although detailed metrics are available on how the Medicare money is being spent across various states, Hospital Referral Regions (HRRs) and counties, only a limited number of tools are available for exploration and visualization of this huge repository of data. HealthCorr is a web-based tool which has been designed specifically for HHS to visualize the Medicare data and other health care data. Specifically, the tool is designed for users who have experience in the health care industry and should be able to achieve the following goals:

1) Enable the users to visualize health performance, quality, and cost indicators for their county and HRR.

2) Enable a user to filter by demographic characteristics (such as age, income, poverty, education). The filters should also hint to the user where their ZIP code of interest lies and the national average lies on the continuum to help the user to make filtering decisions.

3) Spur mayors, doctors, public health professionals and others associated with the health care industry to take action to improve the quality and lower the costs of health care.
Although a few tools exist to visualize the health care data, most of such tools using geospatial visualizations only have the capability to compare a couple of variables at a given time [5,6]. Some other tools have the ability of comparing multidimensional data but lack geospatial abilities [7]. The ability to visualize multiple variables on the map and compare them using a table and charts at the same time can provide valuable insights which might not be possible to obtain from current tools. The novelty of HealthCorr lies in its ability to dynamically depict the geospatial distribution of health data on the map using color coding and to provide detailed information in a table and chart simultaneously, allowing users to compare different regions on the map first for an overview and then go over to the table and chart for more detail. HealthCorr also allows users to discover other geographical regions similar to the one they choose for comparison, in terms of some key demographic indicators, allowing users to make meaningful conclusions about the state of health care relative to similar areas.

The purpose of this paper is to describe HealthCorr in details, and it follows the following structure. In the first section, an overview of the existing tools to visualize health care data, along with their capabilities and limitations is given. Next, key algorithms and variables behind HealthCorr are described in details. In the sections that follow, the interface is described in details, followed by the information architecture of the tool. The paper finally concludes with some insights obtained from the tool, feedback from the clients at HHS and future directions for this tool.

II. RELATED WORK

Data visualization has been used as an important tool to gain insights into health care data sets, which are typically multivariate, discrete and at different granularity levels [Ref]. Previous studies have employed different desktop applications designed for multivariate analysis to study health care datasets [8,9]. However, most of desktop tools for multivariate data visualization typically require the user to spend a significant amount of time learning the tool. Users are expected to have some level of domain-specific knowledge to gain valuable insights [10]. On the other hand, there are a few web applications that exist for health care data visualizations. A majority of these tools use maps to organize and visualize the data, and are designed to be used by laymen having a little or no domain knowledge.

One of the most common tools in this category is the Dartmouth Atlas [11,12]. The atlas features an interactive map of the United States and divides up the regions based on the states and the Hospital Referral Regions (HRRs). It is designed to visualize only two variables, which can be selected by a drop down menu - either the total reimbursements for the year 2006, or the annual growth rate for the selected region from 2001-2006. The map is then colored in five different gradients, depicting the geographical variation for the selected variables. Selecting a particular state or HRR using the cursor also shows the average for the selected region and the corresponding national average.

More advanced version of the Dartmouth Atlas available on the New York Times website [13]. Although the interactive map available on this tool only features the HRRs and not the states, variables from three broad categories - reimbursement, enrollees, and surgery rates are visualized. Like the original Dartmouth, only one variable can be visualized at a given time, and mouse over on a certain region shows a comparison of selected variable value of that region to the state average and national average using bar charts.

One of the web-based tools [14] allows the comparisons of different variables and has been used in several studies is the HHS Hospital Compare. The tool features an interactive wizard which gives the users multiple options to select a geographical region of interest, a single medical condition and returns the most appropriate hospitals based on the input criteria. The tool also allows side by side comparison of up to three hospitals on three selected variables. However, Hospital Compare lacks the ability to compare different geographical regions across the state or the country for different variables, and is only restricted to the comparison of three hospitals in the chosen location.

Although GIS has been found to be a useful method to visualize the health care data [15, 16], there also have been concerns that using interactive maps by itself might be subject to misinterpretations [17]. Our prototype named as HealthCorr gains its inspiration both from the DartMouth Atlas as well as the Hospital Compare. This tool aims to provide the combined features of a simple GIS application.
having the capabilities of comparing multiple variables at any given time and showing the insights using conventional graphs and tables, along with the maps.

III. ALGORITHM

Arguably one of the most challenging aspects of this project is that provided data is in two different geographic resolutions. The data filtering mechanism should allow the user to select counties or HRRs based on specific demographic variables. The filter variables were obtained from census data sets which are compiled at the county level. Many of the interesting health data sets were at the HRR level, which is much bigger than a county and does not respect state, county, or ZIP code boundaries. The problem is further compounded by the fact that selection of a region on the map is performed by searching for a ZIP code, which is smaller than a county and an HRR. The question quickly arose: if the filters variables are the county level and the variables of interest are the HRR level, how does one filter HRRs based on county-level variables? In order to do this, an algorithm was used to compute a population-weighted average of a demographic variable based on the intersections of counties and HRRs. The algorithm is as follows:

For each HRR, run a geospatial query to determine all counties that intersect this HRR

For each intersecting county
- Compute the area of intersection between the county and the HRR
- Compute the total area of the county
- Given the population of county $P_{\text{county}}$,
  \[ P_c = \frac{\text{Area}_{\text{intersection}}}{\text{Area}_{\text{county}}} \times P_{\text{county}} \]
- Sum all of these population portions to get the total HRR population, $P_{\text{tot}}$
- Finally compute the ratio $P_c/P_{\text{tot}}$ and use this as the scaling factor for each filter variable
- HRR value = $\sum(P_c/P_{\text{tot}}) \times \text{CountyLevelValue}$

IV. DESCRIPTION OF THE INTERFACE

The uniqueness of the interface of HealthCorr lies in the fact that it embodies three types of visualizations. The tool simultaneously displays a map, data table, and chart view, allowing the user to see the geospatial as well as multivariate features of the data. The user inputs are provided to the system in three ways: the ZIP code entry panel, selection panel and dynamic filter panel. The visualization dynamically changes according to these inputs.

Fig 1 shows the Graphical User Interface (GUI) of HealthCorr. Essentially, the interface can broadly be divided in four distinct parts. (a) The top left panel consists of a pannable and zoomable interactive map of United States. (b) The bottom left panel consists of a tabbed panel that is used to switch between county and HRR level data tables. The raw data is displayed in a sortable table with cells colored based on the quintiles of the data. (c) The middle left panel contains hinted double sided filters that allow the users to filter the map and data tables simultaneously. (d) The right side vertical selection panel contains a list of variables, classified into four categories, which can be selected using checkboxes.

The right hand side selection panel contains variables that are categorized into four functional areas: Cost, Quality, Accessibility and Public Health. Data for all these variables is derived from the Center for Medicare and Medicaid (CMS) and the Community Health Status Indicator (CHSI) databases. In all, there are 19 variables - 11 at the county level and 8 at the HRR level.

A. Map View

To visualize the data on the map, users begin by selecting a variable to draw the heatmap overlay by clicking the variable name. Once a variable is selected, the link text is bolded and the geographical distribution is shown by heat maps varying in 5 different shades. The four color schemes are derived from ColorBrewer. Users can only select one variable at a time to be shown on the map. Users can zoom (using the map controls or mouse wheel), pan, and obtain the variable value of a region by clicking on the map.

B. Table and Chart view

To view multiple variables, the HealthCorr interface provides two more views: a chart view and a table view. A user can use the checkboxes in the selection panel to check the variables of interest and click the show table link (see Fig 2). The
table shows the values of selected variables within the filter range so that the map and table always stay consistent. Cells are also color coded for easy comparisons with other variables and geographical regions. More intense colors in a cell indicate higher value for a variable (the table colors always mimic the map colors for consistency). Since the data table columns are sortable, the user can easily find the global minimum of a variable by clicking a column heading (and clicking again to reverse the sort direction to find the global maximum). If the user spots an interesting pattern in the data table they have displayed, he or she can create a chart of the values. The show graph link triggers a modal dialog containing a chart to hover above the page while dimming the interface underneath (see Fig 3). The charts are generated dynamically, so the user can create arbitrary charts showing various regions and variables. The charts are particularly convenient for saving a snapshot of an insight or just having one more avenue to perform comparisons.

At this time the table and chart views have some limitations. Some variables are available only at the HRR level, while others are available for the counties. Since these values cannot be compared directly (HRRs do not respect county boundaries), the user can select a tab that contains either a table of HRR variables or county variables. This restriction applies to charts as well for similar reasons. Some ideas have been proposed for how to deal with this limitation (such as showing all counties that intersect an HRR), but have not been implemented in this version.

C. Dynamic Filters

One of the key features of HealthCorr is its ability to employ dynamic filters. Four filtering options represented by double sided sliders are present right below the map. The filters represent the median income, poverty percentage, average age and education level respectively. Using the filters updates both
the map and data table simultaneously by removing entries that are out of the selected range. The double sided sliders are hinted with benchmark values such as the national average and the currently selected region’s value (see Fig 4). By selecting ranges that surround his or her position on the filter, the user can get a sense of which regions are similar to his or her own and immediately see how his or her colors compare to the colors that are remaining on the map and in the table. This is an important feature since a user might try to use socioeconomic factors as excuses to rationalize poor performance or high cost. The ability to filter based on such demographic variables can be helpful in removing the credibility of such excuses. Each time a user picks a new zip code, the your region marker shifts to the proper position on all filter sliders.

### V. APPLICATION EXAMPLE

Having a geographical heat map is extremely effective for showing variables that appear to be clustered around a particular region.

Fig 5 shows an overview of the entire US showing the Average Number of Unhealthy Days in the Last Month. There is a very clear dark spot that develops around eastern Kentucky into West Virginia (circled in red). The counties in this area all fall into the 4th and 5th quintiles and no other place on the map has such a dark area. Initially we were not sure what to make of this – after consulting with a professor at UMD, he pointed out that Kentucky and West Virginia are big mining centers in the US. Indeed, [Wikipedia](https://en.wikipedia.org) confirms that West Virginia and Kentucky have several underground coal mines, which we feel might be plausible explanation of poor health. Furthermore, after zooming into this area and switching the map to life expectancy (Fig 5)
Fig. 3. Table before sorting (top) and table after sorting (bottom). The colors of table cells help the user understand the distribution of a variable and potential correlations.

Fig. 4. Double sided slider for filtering which is annotated with the national average (green dot) and the average for the currently selected region (blue dot).

5), it becomes very clear that the area of poor health corresponds to low life expectancy. Since the map re-renders completely in place, the user can keep their eye on the area of interest to be able to identify correlations with ease. The user can then choose any number of other variable to draw on the map and look for other correlations. While it is somewhat intuitive that coal miners might have poorer health and lower life expectancy, we are encouraged that the tool allows the user to gain these types of insights in a straightforward way.

The map filtering mechanism is also extremely effective in comparing and contrasting different segments of the population. One of the comparisons we explored was to examine different variables for wealthy portions of the country as well as poorer areas of the country (by wealthier we mean median income above national average, poverty rate below national average and by poorer we mean the opposite). Fig 6 shows how wealthier counties report their overall health status and Figure 6b shows how poorer countries their health status. These two maps show the same region of the United States, but we can see that the colors shown on the wealthier map are much lighter colored than the colors shown on the poorer map. This implies that wealthier counties tend to have better health than poorer ones. Again, while this insight may seem to be obvious or trivial,
we feel that it demonstrates the ability of the tool to compare the quality and cost of health care across socioeconomic classes, which is not offered by other tools we have seen.

The filtering mechanism is also useful for trying to pinpoint whether specific demographic variables determine poor quality or high costs. One example can be seen in Fig. 7. This shows an interesting mix of demographic indicators in that it selects regions with high incomes, but also with high poverty rates. It appears that areas such as this have high hospitalization costs for Medicare which can be seen from the dark colors on the map. Currently the tool only has data for two demographic variables (median income and poverty rate), but soon it will have two more (median age and education level).

These four variables should allow the user to dissect the United States in interesting ways to look for patterns that result in high cost or poor quality.

VI. APPLICATION ARCHITECTURE

One of the design goals of HealthCorr is the idea of accessibility. Since the tool broadly targets the people in charge, we do not make any assumptions about technical prowess of these individuals. For this reason we expect that a web application would be the best solution since our target users are not the computer experts. Any user with a modern web browser can view the data with different representations such as a heat-map, tables, and charts, which are generated without any additional plugin, binaries to download, or datasets to import.
Another important design goal of HealthCorr is to make it an *app* in the true sense of the word. HealthCorr does far more than just simply pushing a few tables and mapping the relevant data for users to look at. HealthCorr can be qualified more as a Rich Internet Application in that all interactions happen within a page itself. A user is free to zoom and pan around the map, sort table columns, or pick a new variables with charts to explore all without ever leaving the page. In the past, this type of application was very difficult to build because complex interactions required many components to cooperate. Since this is not a technical report, the technically curious can find a more complete description of the architecture in Appendix A.

VII. INITIAL FEEDBACK

Throughout the project we have worked closely with Todd Park, Karen Milgate, and Patrick Conway and have tried to implement their suggestions since

Fig. 6. Self-reported fair/poor health for wealthy counties of the US (low poverty rate, high median income). (b) Self-reported fair/poor health for the same region, but only considering poorer counties.
Fig. 7. Using filters to select areas with high poverty but also high income, the remaining HRRs appear to have relatively dark colors, indicating that these areas have high hospital costs per capita.

they have a clearer vision of the type of user who would eventually explore data with a tool like this. In our last demo to Todd Park and his team, we have gotten positive feedback on the usability and flexibility of the tool. We wish to share some of the reactions of Todd Park during our latest demos: *This is incredibly cool, I think it’s spare, elegant, and intuitive*, and finally the most encouraging one, *The White House and HHS see this as a powerful seed for changing health care in the United States*. Although we haven’t conducted a formal usability test we feel that based on the reaction of Todd Park and on our own ability to notice some interesting patterns in the data, that this tool has potential.

**VIII. CONCLUSIONS AND FUTURE DIRECTIONS**

HealthCorr has been designed as a simple web application which could enable users to explore the health care data for the geographical regions of their interest, irrespective of their prior experience with other data visualization tool. Some of the key features of HealthCorr are its ability to analyze data at the level of counties, visualize the data simultaneously on a geographical map, charts and tables, dynamically change the values of these variables by selecting a specific range of income and poverty levels of the chosen geographical regions and comparing multiple geographical regions as well as variables side by side. All these features, along with the simplicity make HealthCorr a powerful tool to analyze the health care data in United States and assist in discovering important trends and valuable insights. Our Clients, the Department of Health and Human Services envisions HealthCorr as a tool that would spur action by the decision makers and streamline health care spending. Furthermore, in the long term, HealthCorr could also serve as the centerpiece for a social network of citizens as well as policy makers. Although HealthCorr has primarily been the tool has been designed for the Health Care data, it can easily be used to analyze any multivariate data set with a geographic spread, and could be of particular interest for users in the education or finance domains.

**ACKNOWLEDGEMENTS**

We would like to thank Patrick Conway and Karen Milgate for taking the time to supply us with clean datasets that were easy to use. We would also like to thank Todd Park who provided the initial inspiration for this project and his continued enthusiasm over the tool made it that much more satisfying to develop. Finally we would also like to
thank Dr. Shneiderman for pitching us the idea for this project as well as his continued support.

APPENDIX

APPENDIX A: MORE DETAILED SYSTEM ARCHITECTURE

Web browsers are notorious for being heterogeneous when it comes to their implementations of javascript engines, CSS renderers, and DOM functionality, which can be seen in the compatibility table. In order to ameliorate this situation, we leverage a recent open source javascript library from Google called Closure. This library has an impressive array of UI elements and other advanced features that are used on a daily basis by Google, therefore it is rigorously tested both for correctness and cross-browser compatibility. Other open source projects form the basis of overall architecture of HealthCorr. Geoserver is a powerful map-rendering engine that is committed to strictly following standard sets by the Open Geospatial Consortium. Respecting standards allows Geoserver to easily interoperate with any other application compliant with the standard, such as Google Earth, Microsoft Virtual Earth, OpenLayers, etc. To render the maps generated by Geoserver, we have chosen an open source javascript mapping client called OpenLayers. OpenLayers implement many of the same features as Google Maps, and has a very robust set of features. To generate charts and graphs, we have used an open source PHP library called pChart, which can dynamically draw a plethora of various charts and graphs. All of these tools run on the Tomcat servlet engine (for Geoserver) and the Apache HTTP server. Both of these servers are used extensively all over the Internet. PostGIS is a set of geospatial extensions to the open source relational database PostgreSQL. PostGIS combined with Geoserver form the backbone of all map generation in this project. Finally, to complete the picture, all of this software runs on the GNU/Linux operating system, arguably the most famous open source project of them all. The fact that a group of graduate students working on a project over the course of a handful of weeks can assemble so many independent pieces together into one functioning tool is a testament to just how impressive these open source tools have become. A block diagram showing how these pieces fit together is shown in Fig 8.