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

Paul Rosenfield*, Ginnah Lee*, Awalin Nabila Sopan**, Angela Song-Ie Noh**, and Sohit Karol***
Department of Electrical and Computer Engineering*, Computer Science**, and 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 very trivial and lack the ability to deliver deep insights into this huge repository of data. HealthCorr is a web application designed to visualize the Medicare data in multivariate space as well as geospatially. The tool is designed to be used by an average internet user having a little knowledge about the health care domain. Specifically, 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 and the interface could be updated effortlessly. Healthcorr provides a powerful means for average citizens as well as the policy makers to visualize the current state of Medicare spending, discover peer regions, compare different geographical regions and provide important insights for taking decisions that spur action.

INTRODUCTION

As the health-care debate rages on in Congress, there are sectors of the American public that appear to be somewhat misinformed about health-care. The cost of coverage is exploding for both individuals and businesses, but quality and effectiveness of care are not necessarily following suit. This is further complicated by the fact that costs of health-care at some places are much higher than others, while the quality of health-care is not necessarily related to the cost and there are a large amount of fraudulent claims [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 [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
HealthCorr is a web-based tool which has been designed specifically for the HHS to visualize the Medicare data. Specifically, the tool is designed for users having some experience in the health care industry and aims to achieve the following goals:

1. Enable the users to visualize health performance, quality, and cost indicators for their county or HRR.
2. Enable the users to choose a county or HRR and discover regions similar to theirs, based on some key demographic indicators such as median income, average age, poverty level, and etc.
3. Enable the users to compare their region of interest to other regions in the United States or to the national value as a benchmark for key performance/cost indicators.
4. Spur mayors, doctors, public health professionals and other people associated with the health care industry to take action to improve the quality and 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 having the ability of comparing multidimensional data lacked the geospatial features [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 using the 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 the detailed information on the 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. By using the dynamic filters, users can narrow down the range of some standard variables to see their effect on other variables. HealthCorr also allows users to discover other geographical regions similar to the one they choose for comparison, in terms of some key demographic indicators, which might be a big value for the users.

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.

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 are in 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 require some time to learn by users. Users are expected to
have some level of domain-specific knowledge to gain meaningful 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 used in health-care visualization 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]. The atlas 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 are 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 can be visualized. Like the original Dartmouth atlas, only one variable can be visualized at a given time, and mouse over on 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 that allows the comparisons of different variables and has been used in several studies is the HHS Hospital Compare [14]. 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]. 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 map and dynamic filters.

ALGORITHMS

Arguably one of the most challenging aspects of this project is that provided data is in two different geographic resolutions. For example, a problem we encountered early on was with our filtering mechanism. The filter variables were taken 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 place on the map is performed at the ZIP code level, 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 do we filter the HRR variables? In order to do this we had to approximate the HRR level filter data based on the county level census data. The algorithm computes a population-weighted average based on the areas of intersection between the HRR and counties. 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 the county, compute:
      - \( P_c = \frac{\text{Area of intersection}}{\text{Area of county}} \times \text{Population of county} \)
  - Sum all of these population portions to get the total HRR population, \( P_{tot} \)
  - Finally compute the ratio \( P_c/P_{tot} \) and use this as the scaling factor for each filter variable
    - \( P_c/P_{tot} \times \text{variable} \)
  - Accumulate these partial terms to get the approximate filter value for that HRR

Luckily, PostGIS makes the geospatial computations involved rather easy (although they are very complex to actually compute).

Another algorithm developed for HealthCorr was for estimating a set of peer HRRs given a set of peer counties. The CHSI data set contains sets of peer counties which are separated into 88 stratas. The counties in each set have similar characteristics by age, poverty level, population size, etc. The concept of peer counties is very useful to compare to other counties “like your own”. Since this data is stripped of the actual variables used to compute the peer relationships, we were forced to approximate peer HRR relationships using only this set of peer counties. A somewhat naïve algorithm is used because the problem turns out to be a rather difficult one. The algorithm is as follows:

- For each county in a particular strata
  - Determine which HRRs this county intersects with and compute the intersection area
  - Compute the county's total area and the percentage of the county that intersects that HRR, \( \frac{\text{Area of Intersection}}{\text{Area of County}} \)
    - If the intersection area is <20%, do not add this HRR to the peer list
    - If >20%, add this HRR to the peer list and accumulate the area of intersection

After this algorithm finishes, since counties can cross HRR boundaries, it turns out that a single HRR is likely to end up in multiple stratas. If we take the union of these sets, it includes just about every single HRR, so it is uninteresting. At this point the decision is
Figure 1. The overall interface of HealthCorr which includes the map (top), double sided sliders for filtering (middle), and data table (bottom), ZIP code selection field (top right), the control panel (right) for controlling what is shown on the table and map.

made based on taking the strata with the highest area of intersection. The rationale here is that if the peer counties that defined this peer set had the highest area of intersection with the counties in this strata, then chances are it is the best fit. In reality, this process is extremely fleshy and so the results are most likely not very precise, but it seems to work well enough for selecting a reasonable set of peer HRRs as a demonstration of how such a feature would be implemented and used. In the future, a rigorous statistical analysis of the data should be used to generate the peer HRRs in the same process as the peer counties.

DESCRIPTION OF THE INTERFACE

The uniqueness of this interface of HealthCorr is that it embodies three dimensions of visualization: Map view, Tabular View and Chart view. Figure 1 shows the Graphical User Interface (GUI) for HealthCorr. Essentially, the interface can broadly be divided in three distinct parts: 1. The top left panel of the interface consists of a zoomable interactive map of United States of America (Figure 2) 2. A bottom left panel consists of tabs that be switched between the table and chart view, depicting the raw data and comparison charts for variables of interest respectively (Figure 3) 3. A right side vertical panel, containing a list of variables, classified into four categories,
which can be selected using check boxes for multiple comparisons (Figure 4).

The design of the interface follows the principle of “Overview, Zoom and Details on Demand” [Shneiderman, 19].

**Figure 2:** This figure shows an example of an insight is generated by using the map filtering mechanism. Selecting the richest counties in the US (top, high median income, low poverty rate) produces very light colored counties only (indicating better self-reported health status). Setting the filters to include only low median income and high poverty rate. The resulting counties have a much darker color, indicating that poorer Americans tend to report poorer health.

**Figure 3:** It is possible to sort the table (left is before sorting, right is after sorting on Heart Attack Mortality Rate). The color coding coupled with the ability to sort is surprisingly powerful in that not only does it allow the user to easily find the global min and global max of any variable by sorting, it also gives a sense of the distribution of values. On the sorted table (left) it is apparent that the smallest quintile only has a single value, whereas the next highest quintile has many more values. This signals to the user that this value is an outlier. Another insight that can be observed is that the lighter colored sections of Heart Attack Mortality Rate appear to correspond to darker sections of Medicare Hospital Costs. This signals to the user that higher expenditures on hospitals results in better outcomes for heart attack patients.

**Figure 4:** An example graph that is generated dynamically based on values the user selects. Selecting several HRRs and several variables provides a fast way to compare the given HRRs. Also, saving the image is an efficient way to pass along an insight to someone else.

In the very first view of the application, users are prompted to enter their zip code as soon as they click on the map. The map then zooms to the selected zip code and shows a
marker on the selected area. At any time, users can also enter a different zip code in the field available on the top of the right side panel.

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 level of the counties and 8 at the level of HRRs. Variables depicting the county level data are marked with an asterisk.

Users will select a variable to draw the heatmap overlay by clicking the variable name. Once a variable is selected, its geographical distribution is shown by heat maps varying in 5 different shades. Users can only select one variable at a time to be shown on the map.

We provide two options, to view more than one variables, tabular view and graph view. Once a user select a particular granularity level, only the variables belonging to that level will be available with checkboxes to be selected for table and graph; and all the other variables that do not belong to the same granularity will be available only for drawing the map. In the table display, only the variables that have been selected from the right hand panel show up. Cells are also color coded for easy comparisons with other variables and geographical regions. Further, each of the four categories is represented by a different hue, making it easy to remember what category a particular variable belongs to while doing comparisons. Since the map view only shows the variation of one variable at a given time, the default view of the table shows the variable corresponding to the map. More intensity of color of a cell indicates higher value of a variable and less intensity will indicate smaller value. Storability and gradient color coding for variables provides an understanding of correlation among variables.

One of the key features of HealthCorr is its ability to employ dynamic filters which affects the map view, chart view as well as the table view simultaneously. 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. The national average for these variables is shown by a green dot on the sliders, while the value for the selected region is shown by the blue dots. For any given variable depicted on the map, or a number of variables depicted on the chart/table, data can be narrowed down using the dynamic filters. This filtering option is also good for finding out similar regions.

ARCHITECTURE

One of the key 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 non-Internet Explorer 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. Developers no longer have to worry about how to make the software cross platform or how the user will get new features out of new updated data. We feel that this is a major strength of the tool: the data is
currently on the center of attention and the tool helps the user easily explore it.

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 be cooperated. Javascript is used to issue asynchronous requests to a server (such as a request to draw a new map), the returns data flows into a javascript callback function which updates the page through the document object model interface. Web browsers are notorious for being heterogeneous when it comes to their implementations of javascript engines, CSS

**Figure 5:**

*This figure describes an insight also generated from map filtering mechanism. By selecting the counties with high poverty rate and high median income, we can filter out the counties with small portion of very rich people and high percentage of poor people. When clicking out medicare hospital cost per capita, it shows some counties in Arizona, Tennessee, Ohio and Michigan show the darker red. This shows they are paying unreasonably high Medicare hospital cost per capita although they tend to be poor counties in overall.*
renderers, and DOM functionality, which can be seen in ugly compatibility tables such as [http://www.webdevout.net/browser-support-summary]. 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 [http://geoserver.org/display/GEOS/What+is+Geoserver]. 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 [http://openlayers.org]. 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 [http://pchart.sourceforge.net]. All of these tools run on the Tomcat servlet engine (for Geoserver) and the Apache HTTP server [http://www.apache.org]. Both of these servers are used extensively all over the Internet. PostGIS [http://postgis.refractions.net] is a set of geospatial extensions to the open source relational database PostgreSQL [http://www.postgresql.org]. 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 figure 5.

**USABILITY STUDY**

In lieu of a formal usability study, we simply wish to share the reactions of Todd Park who is our main customer during our numerous 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". Perhaps the best news is that these reactions were in response to only a partially completed tool.

**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 any data visualization tool. Some of the key features of HealthCorr are its ability to analyse 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. Further, in the long term, HealthCorr could also serve as the centerstage 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.

ACKNOWLEDGMENTS
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.

References


[10] Shneideran – MILC


