Creating a Geographic Health Information System to Analyze Spatial and Social Patterns of Emergency Department Usage in Olmsted County, Minnesota, USA

Joshua J. Pankratz[1,2]
[1]*Department of Resource Analysis, Saint Mary's University of Minnesota, Minneapolis, MN 55402 ;* [2]*Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905*

Keywords: GIS, Emergency Department, Health Research, Rochester Epidemiology Project, Socioeconomic Status, Olmsted County, Minnesota

Abstract

Techniques used to geographically visualize patterns of health have been around for a long time, but even today many modern medical institutions do not regularly utilize geographic analysis methods to learn patterns of health usage within their practices. Geography and health information can be combined to produce a valuable infrastructure for new types of research and knowledge discovery. The goal of this research was to create and utilize a geographic health information system (GHIS) to investigate spatial and social patterns of Emergency Department (ED) usage in the population of Olmsted County, Minnesota, USA residents. This was completed by (a) creating a GHIS that combined patient electronic medical record (EMR) data and publically available indicators of socioeconomic status and lifestyle attributes from national survey data sources, (b) using various spatial analysis techniques to investigate clustering of patients who utilize the ED on a frequent basis, and (c) using statistical analysis techniques to find social or demographic characteristics correlating with patients who are high ED utilizers. There were statistically significant geographic clusters of patients with high ED utilization along with statistically significant demographic factors, such as median income, that are highly correlated with ED utilization rates. Findings demonstrate using a combined set of geographic and EMR data, an infrastructure can be built to answer important questions and find patterns that may otherwise go unnoticed.

Introduction

The topic of Emergency Department (ED) utilization is something many medical institutions analyze closely. ED care is very expensive and meant to be used for acute and severe conditions. Overuse of EDs for lesser priority conditions that could be handled in another setting increases costs and congestion of EDs.

According to LaCalle and Rabin (2010), hospital ED patient volumes have increased 36% from 1996 to 2006. This has prompted many providers to evaluate underlying reasons why patients utilize emergency departments and help to implement interventions if other methods of care can more effectively treat patients' needs.

Traditionally there is not much socioeconomic information collected in electronic medical record systems even though socioeconomic status may be a significant factor in a person's health. This information can be valuable in helping to describe underlying factors that influence a patient's pattern of healthcare. According to Miranda, Ferranti, Strauss,

Pankratz, Joshua J. 2014. Creating a Geographic Health Information System to Analyze Spatial and Social Patterns of Emergency Department Usage in Olmsted County, Minnesota, USA. Volume 16, Papers in Resource Analysis. 16 pp. Saint Mary's University of Minnesota University Central Services Press. Winona, MN. Retrieved (date) http://www.gis.smumn.edu

Neelon, and Califf (2013), EMR systems are typically lacking analytical tools required to connect disparate sets of data from non-health care sources such as social and environmental information. Inclusion of patient location along with geographic information system (GIS) analyses and area-based socioeconomic measures can enhance the information available to better understand patients. Research with this work show how using GIS tools and techniques can create an enhanced data infrastructure enabling answers to questions like finding spatial or social patterns of ED usage for a patient population.

Significance

If geographic patterns or socioeconomic patterns are shown to be different in patients who utilize ED most often, policies and procedures can be developed to ensure optimal care can be delivered to those patients. Miranda *et al*. (2013) asserts if hospitals can understand key drivers and spatial patterns of ED usage, an improved experience and better health care outcome can be achieved by directing patients to a more appropriate setting, along with reducing cost.

Another important topic to address is disparities in the way health care is delivered between different demographic segments of the population. These disparities are typically defined as a "systematic difference in the use or receipt of health care services between white and nonwhite individuals, people with and without disabilities, rural versus non-rural dwellers, or people with high versus low education, who have comparable need for services" (Begley, Basu, Lairson, Reynolds, Dubinsky, Newmark, Barnwell, Hauser, and Hesdorffer, 2011). However, the underlying infrastructure to answer these types of questions is largely not present in most medical institutions. This research emphasizes the methodology of creating a GHIS to link standard medical data and geographically referenced data. This type of infrastructure can be utilized to answer various questions beyond those discussed in this paper.

Methods

Research Population

The cohort of study was the patient population of Olmsted County, Minnesota. The Rochester Epidemiology Project (REP) is a medical research infrastructure which has followed the population of Olmsted County for the past 48 years (St. Sauver, Grossardt, Yawn, Melton, and Rocca, 2011). The REP has linked the residents of Olmsted County with the various medical providers they utilize for their health care needs and also provides linkages for researchers to the data associated with those patient health records. There are two emergency department locations in Olmsted County (Figure 1).

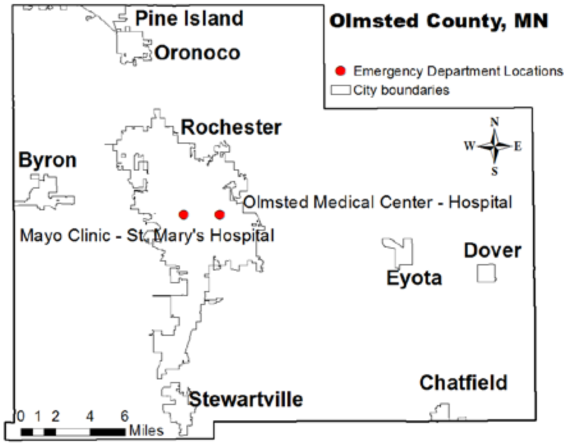

Figure 1. Medical emergency department locations in Olmsted County, Minnesota shown relative to the cities within the county.

The REP medical records linkage system was used to identify all residents of Olmsted County, Minnesota on April 1, 2009. Only patients who gave permission for review of their medical records for research (147,377 patients) were included. This cohort was defined by St. Sauver *et al*. (2013). Details of this REP census infrastructure are described in St. Sauver *et al*. (2012).

In order to utilize this patient population, and use patient data in this work, Dr. St. Sauver provided permission for the use of this study's data and facilitated approval processes for this project to get institutional review board (IRB) approval from the Mayo Clinic IRB and Olmsted Medical Center IRB.

The next step was to re-link patients to their medical records based on current information and re-screen them to verify their research participation consent. This process removed 4.38% of the patients, leaving 140,924 patients remaining in the cohort. Patient addresses then had to be translated into latitude and longitude points in order to utilize their location data in the GIS analyses. End results reduced the total cohort to 130,678 patients with valid addresses.

After defining the cohort, vists to any Olmsted County ED were collected and summarized. The duration chosen for this analysis was one year – from January 1, 2008 to December 31, 2008. A query was performed to find data on all ED visits. Both the Mayo Clinic and Olmsted Medical Center have various indicators in their data to flag if a visit was initiated from the ED. Those flags were used in the query to obtain a fully representative list of ED visits.

After compiling ED visits for each patient, a final sum of visits for each patient within the one-year timeframe was calculated and saved to be used for analysis.

To focus on patients with high ED utilization, a threshold was chosen to define high utilization. In investigations by LaCalle and Rabin (2010), they found a variation in definition across various research studies that defined high utilization anywhere between 3 visits per year and 12 visits per year for a single patient. The definition of frequent ED usage was typically dependent on the goals of research. The most agreed upon metric from these previous studies was ≥4 visits to the ED within one year, with the rationalization being that ≥4 visits represented 25% of all ED visits, which was substantial enough to warrant investigation.

In the Olmsted County cohort, the data were significantly different in regards to the percentage of patients visiting the ED. Patients with ≥3 ED visits during 2008 only represented 3% of the 2009 Olmsted population (Table 1). Since the ED utilization rates for Olmsted County appear different than other research publications, a sensitivity analysis was used. For spatial models, values of ≥3 visits to ED were chosen as the definition for this project as the "high utilization" threshold, but statistical analyses in this research performed several definitions of high utilization for testing and comparison.

Table 1. A frequency distribution showing the number of Olmsted County residents classified into the number of ED visits they had in the year 2008.

# ED visits	Frequency	Percent
0	104,759	80.17
1	16,911	12.94
2	5,220	3.99
3	1,859	1.42
4	822	0.63
≥5	1,107	0.85

Using Area Based Socioeconomic Measures

In standard electronic medical record systems, information such as income level, occupation, education, and other socioeconomic type information are not consistently collected (Begley *et al*., 2011). This makes it difficult to perform research on patient populations with certain demographic characteristics or ascertain patterns in health care utilization by these attributes.

Many research projects address this limitation by conducting specific surveys of their patient population, such as LaCalle and Rabin (2010) and Hunt, Weber, Showstack, Colby, and Callaham (2006). Collection of this data via survey can introduce additional biases and potential data analysis problems. For example, according to Beebe, Ziegenfuss, St. Sauver, Jenkins, Haas, Davern, and Talley (2011), populations of patient groups who do or do not respond to surveys have age and race differences. This may introduce bias and potentially skew the results. Their recommendation is, when possible, to link to the U.S. Census as a method of comparing and gauging the level of bias that could be introduced with a survey, or to use the U.S. Census information as a non-biased set of data that can be evenly applied to the entire population.

U.S. Census survey datasets can be retrieved at a variety of different geographic levels (Figure 2). As this project was evaluating a specific county, the geographic levels scaled below county geographies consist of: tracts, block groups, and blocks. According to Krieger, Chen, Waterman, Soobader, Subramanian, and Carson (2003), the level of geography mattered in their study results. Their findings indicated block groups or tracts were geographic levels easiest to link their patient population to and also provide meaningful socioeconomic measures. Olmsted County has 33 tracts and 111 block groups. The block group level was chosen for this project because it provides more granularity than tracts, which were very broad areas within the county.

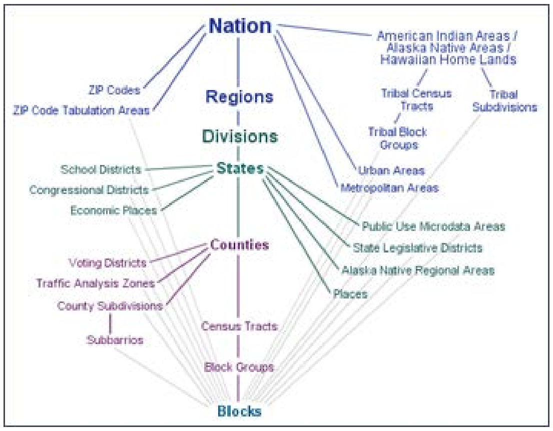

Figure 2. Standard Hierarchy of Census Geographic Entities, U.S. Census Bureau.

Because the timeframe for this study was evaluating the patient population in the year 2009, the American Community Survey (ACS) 5-year data, which spanned 2007-2011, was used to represent demographic characteristics of the patient population. This information was downloaded from the U.S. Census website.

There were five data variables included from the ACS dataset at the block group level: median age, median annual household income, percentage of the population who are white, percentage of the population who are unemployed, and population density of the block group. These variables were defined in the ACS dataset with the following variable names, or were derived by the following formulas (Table 2). A final dataset was created for the 111 Olmsted County block groups with these attributes included for further analysis.

Table 2. ACS variables used in this project.

Data element	ACS variable
Median age	B01002e1
Median annual household income	B19013e1
Percentage white	B02001e2/B02001e1 (white population divided by total population)
Percentage unemployed, age 16+	B23025e5/B23025e1 (population unemployed divided by total population)
Population density of the block group	B01001e1/ALAND10 (population divided by land area)

Joining Spatial and Non-Spatial Data

In order to effectively utilize ACS data, it was joined to the patients in the study cohort at the individual level (Patel and Waters, 2012). The first step in this process was to obtain the most relevant address from the timeframe of interest and translate that address to a latitude and longitude location, a process called geocoding. The addresses were retrieved by running a SAS retrieval "macro" which returned the entire historical set of addresses for each patient. Attached to each address was the date it was collected; the address for the date nearest to 7/1/2008 was chosen for each patient.

Addresses were obtained from the respective medical institutions where they were hand-entered by their staff. Upon reviewing the data, it was determined addresses were not well standardized. In order to perform more accurate geocoding, addresses were processed through an internal Mayo Clinic web service that performs spelling correction, address standardization, and address validation.

Next, patient information and address information was converted into a .dbf file in order to import it into the ArcGIS desktop software application. To obtain accurate geocoding locations, first a local dataset of Olmsted County address centerlines were purchased from the Olmsted County Planning and Zoning Department. ArcGIS U.S Maps and Streets data were used as a secondary source for geocoding reference. A composite address locator was created to first look at the local Olmsted County reference data, then as a fallback the national U.S. reference data to determine a proper location for each address.

After performing geocoding processing on patients who had valid addresses, 90.73% of addresses matched to a location, and 9.27% were unmatched. The final number of patients in the study cohort who had accurate geocoded information was reduced to 130,678.

The ACS shapefile with the 111 Olmsted County block groups was then loaded into ArcGIS and a spatial join was performed between the polygon layer of block groups and the point layer of patient locations. This process enabled the creation of a polygon block group shapefile which contained a sum of the total number of patients in each block group and the sum of the total number of patients who qualified as being high ED users within each block group. This allowed for spatial analyses to be performed at the block group level. Another spatial join was performed in an alternate direction where the block group shapefile was joined to the patient point file, indicating for each patient which block group they resided in. This allowed additional specific analyses at the individual level to be performed.

Spatial Analysis and Mapping Methods

To obtain an overall sense of where patients lived that utilized ED services, several spatial analysis techniques were performed to create maps showing patterns of usage across Olmsted County. Common to all mapping approaches was the use of ≥3 ED visits within one year as the benchmark for classifying a patient as a “high utilizer” of ED services.

Number of High Utilizers

The first approach was to represent the count of patients classified as high ED utilizers, grouped at the U.S. census block group level. This method portrays a representation of the total number of patients who fell into the category of a high ED utilizer across Olmsted County.

Kernel Density Incidence Map

To obtain a distribution of patients with high ED utilization, a kernel density estimation was performed. This was done by creating a point shapefile which contained a point for every patient who had ≥3 visits to the ED. The Kernel Density analysis tool was run with varied parameters, using a search radius from 500 meters to 4000 meters, and using the recommended cell size of 526 meters.

High Utilization Rate

There are many ways to visualize high utilization rates across Olmsted County. This project took two different approaches to visualize this data. The first approach involved calculating the rate of patients who were classified as high ED utilizers based on counting these patients within census-defined block groups, and dividing that value by the total number of patients who live within that block group. The total patient denominator value was calculated based on the REP patient denominator, not the U.S. Census denominator to keep a consistent scale with the patient data that was examined.

The second approach was to create a fishnet grid of 500 meter squares across Olmsted County. Then, the number of patients with high ED utilization was divided by the total population who live inside each square. This was possible because the denominator of the entire 2009 Olmsted County cohort exists within a point shapefile and geocoded to a specific location. This provides a way to calculate and visualize rates with uniform geographic sized area partitions instead of varied size block groups.

Hot Spot Analysis (Getis-Ord Gi*)

The spatial analysis method that seems to be most applicable in this scenario was the Getis-Ord Gi* hot spot analysis technique. This returns statistically significant clusters of patients with high ED utilization per population.

Determining the correct input parameters to conduct a hot spot analysis was a non-trivial task. Two methods were used and compared. The first methodology used was a conceptualization of spatial relationships called zone of indifference. This allowed for a fixed distance for block group neighbors to be set and then used all other neighbors in a fading level of impact the farther away that neighbor is. An incremental spatial autocorrelation analysis was conducted to find the best distance measure to ensure the maximal clustered effect on the data. The results of that incremental spatial autocorrelation showed that the z-score that had the most impact was at 2,500 meters distance (Figure 3). This value was then used in the hot spot analysis tool.

The second type of hot spot methodology was changing the conceptualization of spatial relationships

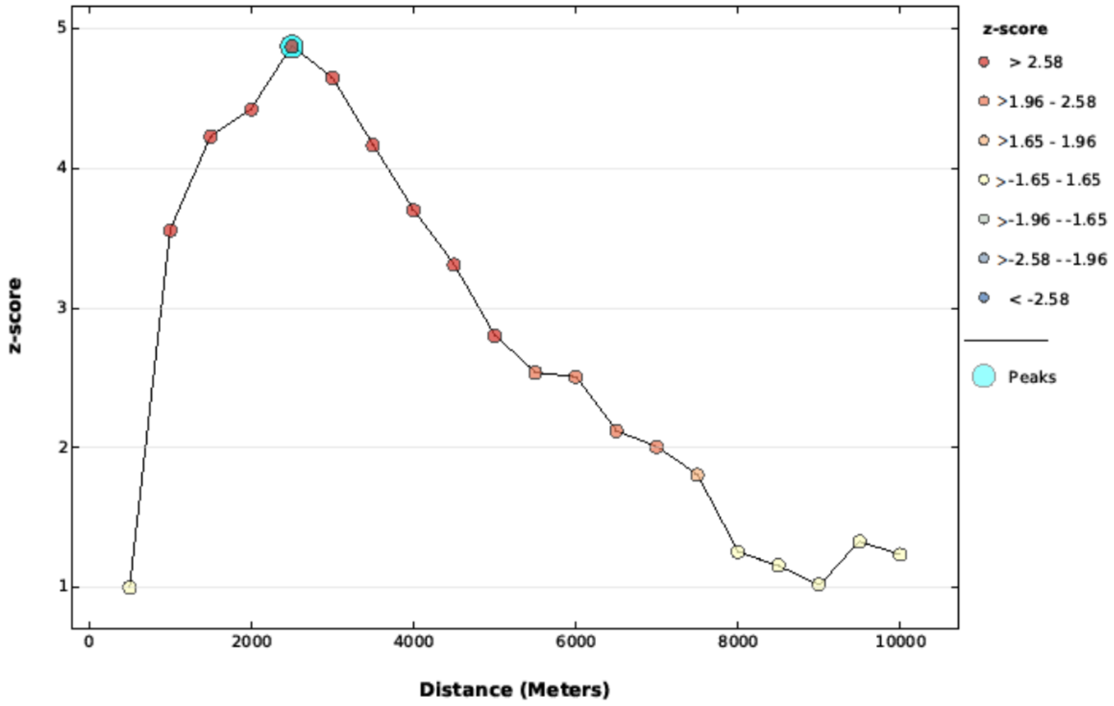

Figure 3. Incremental spatial autocorrelation z-score results, used to find optimal distance threshold.

to utilize polygon continuity on edges and corners. Only block groups with adjacent edges or corners were used in this analysis. The impact of this change was the scale of the analysis changes based on the size of the block group polygons. The larger the polygons, the larger the spread of analysis; the smaller the polygons the more focused the scale of analysis. This method of hot spot analysis was used to provide a contrast to the zone of indifference analysis and see what impact the parameter changes had on the overall outcome.

Statistical Analysis

In order to explore more in depth how different socioeconomic factors may be related to utilization of ED services, a set of statistical models were run to assess patterns within the data. The rate of Olmsted County residents who were classified as high utilizers of ED services within the population was the dependent variable that was used in the regression models.

When evaluating this rate data to determine an appropriate regression model, it was clear the data was not normally distributed. The most appropriate model was determined to be a negative binomial rate regression. Five variables were used as independent variables: median age, median annual household income, percent white, percent unemployed, and population density in units of people per km^2.

To understand results of the regression model, the data needed to be classified in such a way that a unit change in the data would be a meaningful change. The data was organized in this manner:

1) Median annual household income, units of $10,000
2) Population Density, units of 100 people/km^2

3) Percent white, increments of 10%
4) Age, increments of 10 years
5) Percent unemployment, left as is at single units of percent

Results of this regression model were such that a unit change in the independent variable corresponded to a percent change in the rate of patients classified as high ED utilizers.

First, univariable models were examined for each of the five independent variables individually. Multivariable models were used to consider all variables simultaneously. In the sensitivity analysis, the criterion used to classify patient as a high utilizer of ED services was varied. Definitions of high ED utilization of ≥1, ≥2, ≥3, ≥4, and ≥5 ED visits were modeled.

The general equation for negative binomial rate regression models looks like this:

$$\log(\mu) = \log(n) + x'\beta$$

In this equation, μ is the mean event count, n is the total denominator of the rate, and x is the various predictor variables. In SAS, the model was fit as follows, using percent white as the predictor variable as an example:

```
proc genmod data=analyze;
  model count = Perc_White10
  /dist=negbin link=log offset=ln_denom;
  output out=_pnb p=pred_nb;
run;
```

Where:

1. count = the number of patients classified as high ED utilizers
2. Perc_white10 = the percentage of the population who are white in 10% increments
3. ln_denom = the log of the population denominator

This code was run for each of the five independent variables. Similar code was used for a multivariable model including all five independent variables together.

Results

Spatial Analyses and Visualizations

The count of patients with high ED utilization, grouped at the census block level, was mapped to show an overall picture of the number of patients with high ED utilization across Olmsted County (Figure 4).

The next representation of this data was to perform a kernel density estimation on the individual patient locations across the county. This type of analysis gives a sense of the number of patients with high ED utilization in Olmsted County, but not necessarily aggregated by a Census boundary. In Figure 5, the smoothed kernel density representation is illustrated. These initial two representations of the number of patients with high ED utilization will also be, as a side effect, highlighting areas with higher populations across the county. As population density increases, there will be more patients who fall into that category.

In order to account for varying population density across Olmsted County, two different types of output graphics were created to display rates of patients classified as high utilizers of ED services per population. These two approaches produced a consistent measure of how many patients are high ED utilizers per thousand of the population.

The first method represented was by the high utilizer rate data by census block group boundaries (Figure 6). In Olmsted County, there are some block

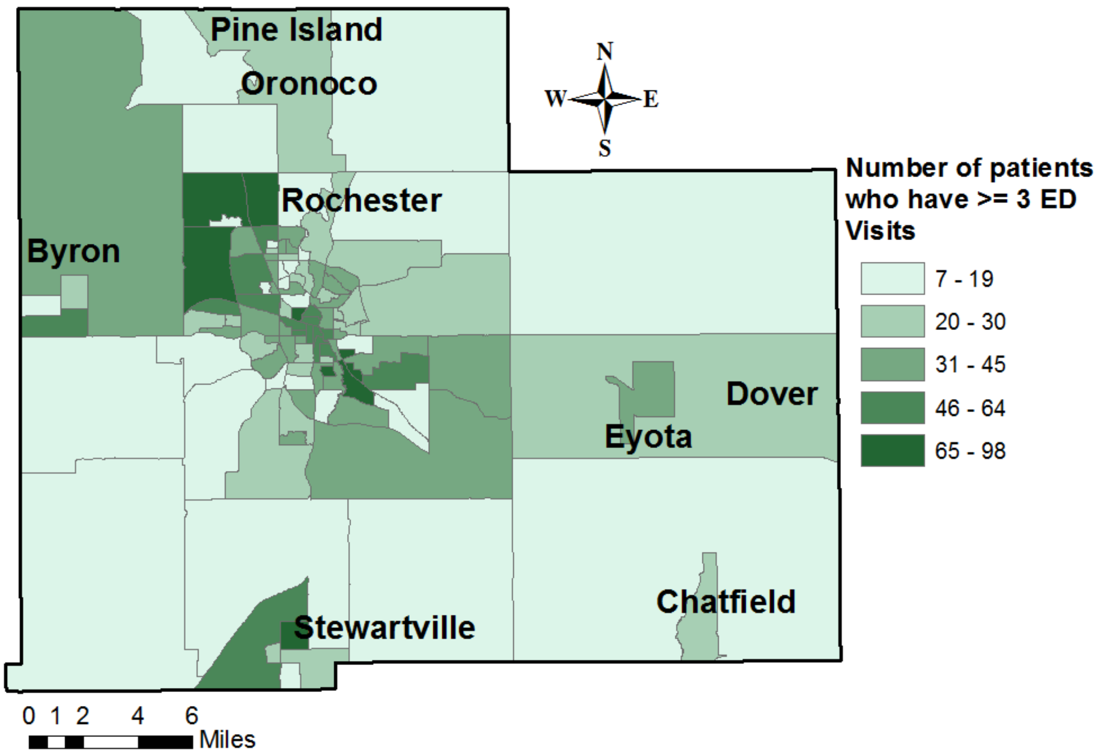

Figure 4. Number of patients with high ED utilization per block group in Olmsted County, MN, classified using natural breaks.

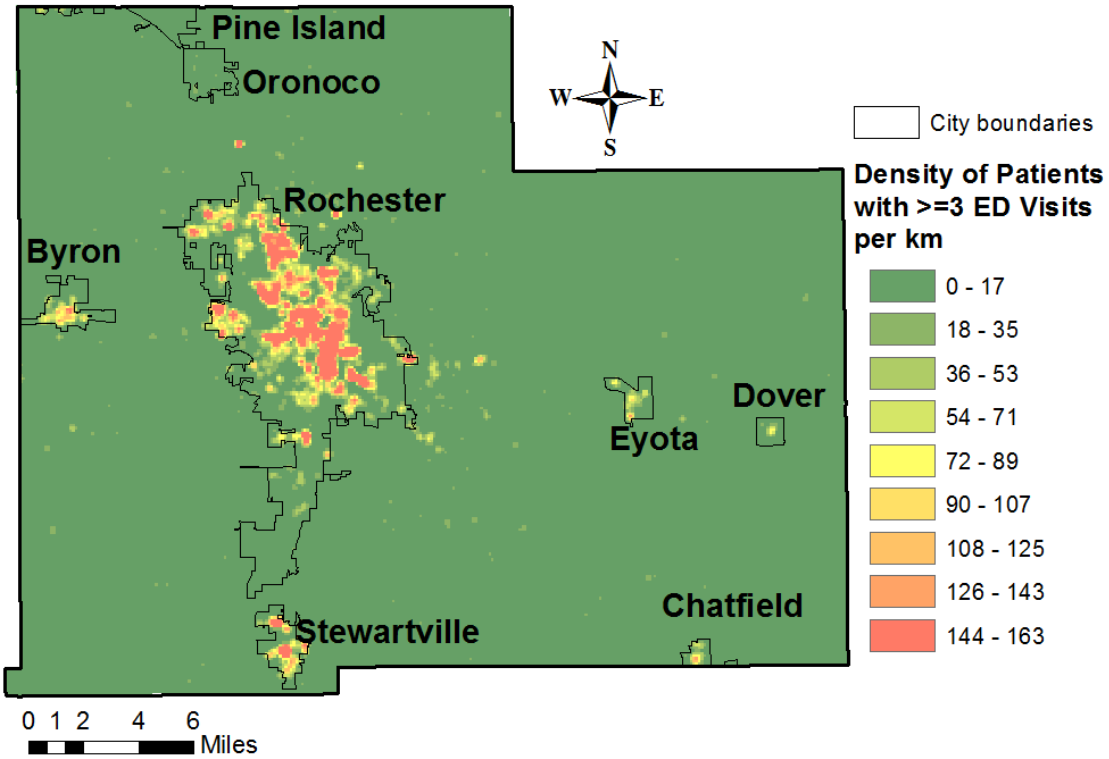

Figure 5. Kernel density map of patients with high ED utilization per kilometer in Olmsted County, MN.

groups covering much larger areas than others. This can lead to patterns of high utilization rates getting lost in large aggregated geographic areas. To assess this in a different way, another high utilizer rate map was created, based on a fishnet grid of 500 meters across Olmsted County (Figure 7). This allowed for an equal divide of geographic areas from which high utilizer rates per population could be calculated.

The final set of spatial analyses that were performed involved creating hot spot maps to find clusters of similarly high or low rates of ED utilization. The first hot spot map was created using a zone of indifference polygon relationship setting (Figure 8). The distance parameter used was 2,500 meters, determined as a result of the incremental spatial autocorrelation analysis. The value of 2,500 meters was the peak z-score, indicating the optimal distance for maximum clustering.

The second hot spot analysis was created using polygon continuity on edges and corners relationship setting. This changes the way neighboring block groups affect each other in the hot spot analysis. Results of the second hot spot map (Figure 9) showed some subtle differences in high and low clustering than the first hot spot map.

Statistical Analysis Results

Different negative binomial regression models were run to compare the rate of high ED utilization for the five different socioeconomic indicators. In addition, the definition of a patient with high ED utilization was varied and the results compared in a sensitivity analysis.

Table 3 shows univariable results for varying definitions of high ED utilization. Beta estimate values and p values are reported separately for each definition of high ED utilization (separate columns). Results can be compared across columns to look for consistency in results for varying definition of high ED utilization.

To interpret meaning of this table, the beta estimate shows the percentage increase or decrease of the rate of patients who are high ED utilizers on a unit change of the independent variable. For example, the independent variable of median annual household income, in $10,000 increments, was shown to be statistically significant with a p value of 0.001 when high ED utilization was defined as ≥1 ED visits. The resulting beta value was -0.063. When put into a narrative, this means that for every unit ($10,000) increase in median annual household income, there is a 6.3% decrease in the rate of people who are high ED utilizers. Results were consistent across definitions of high ED utilization with a trend for greater decrease in rate of high ED utilization as the number of visits used to define utilization increased.
The only independent variable that was not statistically significant for any analysis performed was the median age of the block group. Each test performed produced a p value much greater than 0.05. In addition, the only variable that lost its significance as the definition of high ED utilization increased was the percent unemployment, which was marginally significant (p=0.08) with a definition of ≥5 visits for high ED utilization.

In the multivariable negative binomial rate regression model, considering all variable simultaneously, the only significant variable was median annual household income (Table 4). All other variables had a p value >0.05. Results were similar for all five definitions of high ED utilization. In addition, since median age was not found to be significant

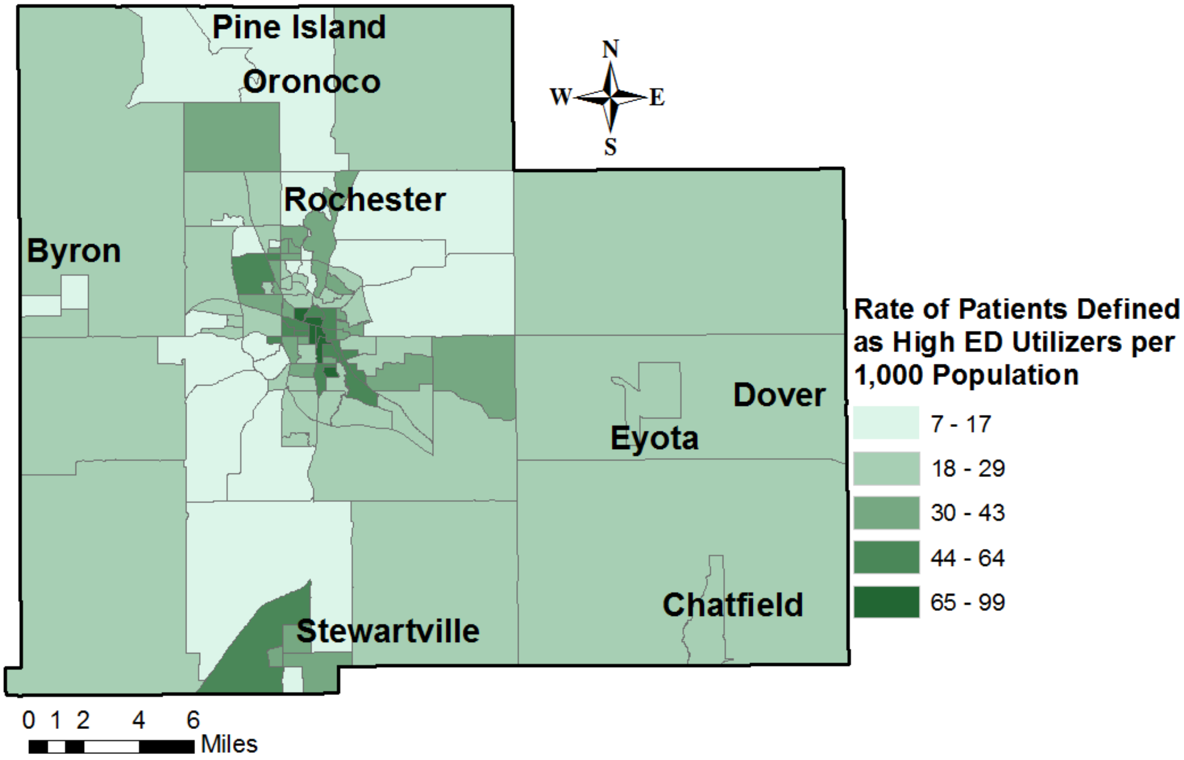

Figure 6. High ED utilization rate by block group in Olmsted County, MN, classified using natural breaks.

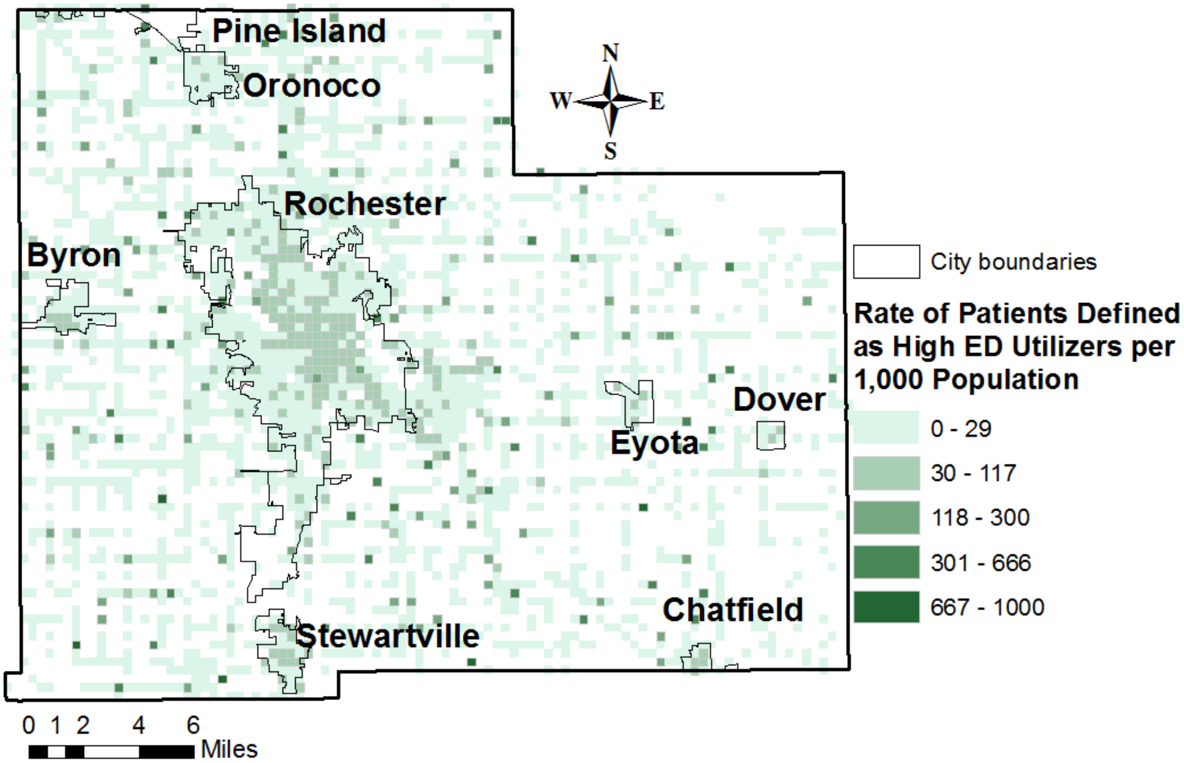

Figure 7. Rate of patients classified as high utilizers of ED services (≥3 visits) per 1,000 population, by 500 meter grid in Olmsted County, MN. Rate value ranges are classified using natural breaks.

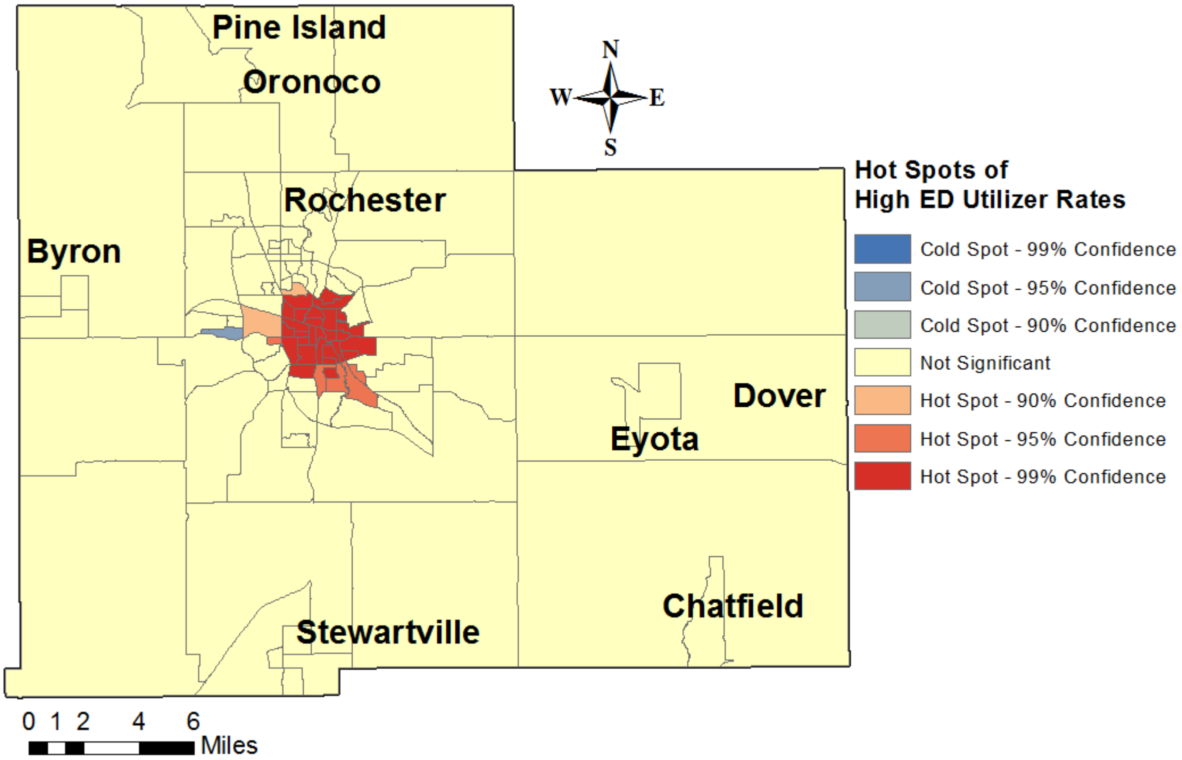
Pine Island
Oronoco
Rochester
Byron
Dover
Eyota
Chatfield
Stewartville
N
W
E
S
Hot Spots of
High ED Utilizer Rates
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence
0 1 2 4 6
Miles

Figure 8. Hot spot analysis of high ED utilizer rate in Olmsted County, MN, using zone of indifference at 2,500 meter distance threshold.

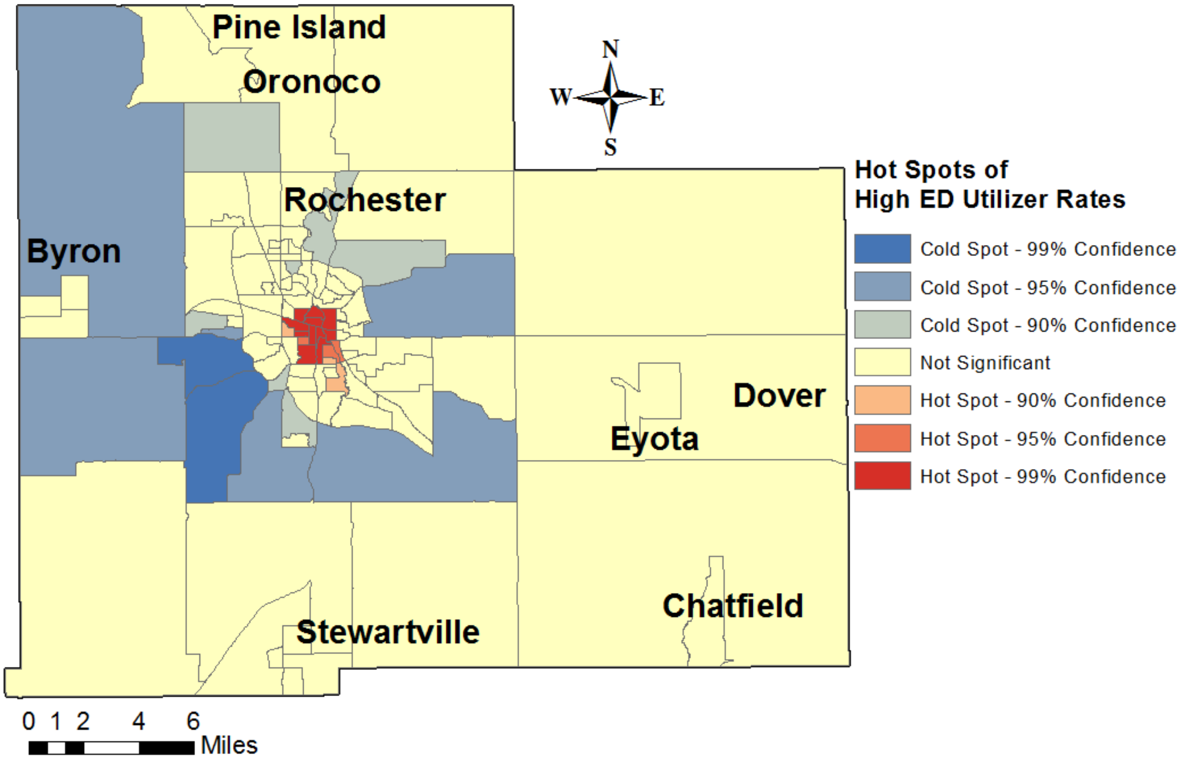
Pine Island
Oronoco
Rochester
Byron
Dover
Eyota
Chatfield
Stewartville
N
W
E
S
Hot Spots of
High ED Utilizer Rates
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence
0 1 2 4 6
Miles

Figure 9. Hot spot analysis of high ED utilizer rate in Olmsted County, MN, using polygon continuity.

Table 3. Results from the univariate negative binomial rate regression.

High Utilization Definition	**≥1 ED Visits**		**≥2 ED Visits**		**≥3 ED Visits**		**≥4 ED Visits**		**≥5 ED Visits**	
Variable	β	*p*	Β	*p*	β	*p*	β	*p*	β	*p*
Median age (10-yr increments)	0.012	0.730	0.046	0.416	0.088	0.238	0.117	0.195	0.133	0.275
Median annual household income ($10k increments)	-0.063	0.001	-0.108	0.001	-0.145	0.001	-0.180	0.001	-0.205	0.001
Percentage white (10% increments)	-0.062	0.001	-0.115	0.001	-0.154	0.001	-0.170	0.001	-0.199	0.004
Percentage unemployed, age 16+	0.023	0.001	0.036	0.002	0.042	0.007	0.042	0.031	0.046	0.082
Population density of the block group (100 people/km^2 increments)	0.011	0.001	0.020	0.001	0.028	0.001	0.035	0.001	0.040	0.001

in the univariable models, the multivariable model was also tested excluding age. In addition, the multivariable model tested various combinations of 2, 3, or 4 independent variables as well. The end result was the same, with only median annual income found to be significant.

Discussion

Spatial Analyses

The maps included in this paper show definite patterns of high ED utilizer rates within Olmsted County. There are statistically significant areas within the center of the city of Rochester that warrant closer attention. These particular areas are prone to seeing higher rates of people coming into the ED on a more frequent

Table 4. Multivariable negative binomial rate regression model results, using all five independent variables. The high ED utilization definition was ≥1 visits.

Variable	**β**	***p***
Median age (per 10-yr)	0.015	0.478
Median annual income (per $10,000)	-0.058	<0.001
Percent white (per 10%)	-0.014	0.217
Percent unemployed, age 16+ (per 1%)	0.003	0.509
Population density of the block group (100 people/km^2)	0.002	0.265

basis. The hot spot analysis map, even using two different methods of analysis, show a pronounced area of high utilizer rates in the core 11 block groups in central Rochester. The hot spot analysis was run two different ways specifically because the block group sizes in Olmsted County vary greatly in both population and geographic size and this can affect the clustering analysis. If the analysis uses straight fixed distance, then the largest of the block groups may not reach enough neighbors to create a proper analysis, whereas the same fixed distance could smother more detailed cluster patterns in the very small block groups within central Rochester.

The zone of indifference is a good mix of using a fixed distance, but then also including additional neighboring block groups outside of that distance but at a less significant level. Also, the polygon continuity method is a way to have a consistent number of neighbors for an analysis regardless of what size the block groups are; as an artifact of that, scales of a hot spot analysis changes based on block group size. Both methods that were run produced valid results. Subtle differences in results between the two are explained by the different purposes of running the two clustering methods.

When looking at the high ED utilizer rate – grouped by fishnet grid, there are very high rates in certain outlying areas that seem out of place. Upon review, the fishnet grid size of 500 meters was probably too small of an area because a grid area with very few people could have artificially high rates. That situation can occur when a small number of people fall into that high utilization category but where there is also a very small population denominator within the grid cells. For next steps, enlarging the fishnet grid size would help eliminate some of those high rate artifacts on the map and make the analysis a bit more useful. However, this fishnet grid map proves block groups are not the only means to aggregate rate information as the study cohort includes all residents of the county and they can then be aggregated at any geographic level or size that makes sense for the type of analysis.

Statistical Analyses

There are significant clusters of areas with high rates of ED utilization. Statistical analysis attempts to help identify underlying factors that may contribute or correlate to that utilization. As Table 3 shows, the variables median annual household income, percent population white, percent population unemployed, and population density were all significant when tested individually with the rate of high ED utilizers. These results were seen consistently across definitions of high ED utilization. With high ED utilization defined as ≥5 visits, the rate of patients who are high utilizers decreased by 20% with a unit increase in median annual household income or percentage of white population.

In multivariable models, considering all variables simultaneously, only median annual household income remains significant. The most reasonable explanation of this result was that several of the predictor variables may be highly related to each other and may not add unique or independent information into the model. Median annual household income is closely related to percentage unemployment, and may also be related to the percentage of white population. If several of these variables are closely related to each other, then a single variable may adequately explain the association between socioeconomic status and rate of high ED utilization.

Limitations and Future Direction

There are several aspects of ED utilization rates that were not explored in this paper. One was retrieving and classifying the reasons for these ED visits. This could shed some additional light on exactly what types of visits people were coming in for, and if the types of visits differ between the patients who do and do not frequent the ED.

Another factor to investigate would be additional types of demographic variables from the ACS above and beyond the five used in this project's analyses. There are dozens of topics addressed in the ACS, and those data are translated into hundreds of interesting variables. Further analyses could explore these other interesting factors.

Another potential limitation was not being able to fully investigate the reasons for why a subset of the cohort could not be fully geocoded. There could be factors at play as to why they did not have a geocodable address that could influence the end result of the analyses. Homeless populations, people staying in temporary housing, or a rural population using post office boxes are all potential reasons as to why 9% of the cohort had either no address or an address that was unable to be geocoded. More investigation could be conducted to identify potential patterns or bias in that set of patients.

Conclusions

The goal of this research was to combine geographic information, census survey information, and patient EMR information to create a geographic health information system which will allow new types of questions to be answered and new knowledge to be gained.

The specific question that was investigated was to see if there were spatial or socioeconomic patterns associated with the rate of emergency department usage in a cohort of Olmsted County residents. The final analyses showed there were statistically significant clusters of block groups with patients defined as high ED utilizers, indicating a noticeable spatial pattern to ED usage. Five socioeconomic variables were explored in a multivariable regression model, with median annual household income showing statistical significance. The strong relationship discovered between lower median income and higher rate of ED usage could be an indication of populations with greater SES having easier access to primary care resources, perhaps better insured, and having less acute need for ED services. This knowledge can direct further investigation into how to better serve the population more at risk for high ED utilization.

This paper shows a proof of concept suggesting integrating geographic data with medical data can open up a whole new perspective when researching health-related questions. There are many more paths that could be taken to expand this investigation.

Acknowledgements

I want to thank my colleagues on the Rochester Epidemiology Project for their invaluable advice and guidance. I want to thank Dr. Jennifer St. Sauver for allowing me to extend the work she had done on the 2009 Olmsted County population.

I would also like to thank Mr. John Ebert for his guidance and for creating such a positive learning environment at Saint Mary's University over the course of my studies.

Finally, this project would not have been possible if not for the valuable research data infrastructure provided by the Rochester Epidemiology Project grant. That grant is supported by the National Institute on Aging of the National Institutes of Health under award number R01 AG034676.

References

Beebe, T., Ziegenfuss, J., St. Sauver, J., Jenkins, S., Haas, L., Davern, M., and Talley, N. 2011. HIPAA Authorization and Survey Nonresponse Bias. *Med Care*, *49*(4), 365-370. doi:10.1097/MLR.0b013e318202ada0

Begley, C., Basu, R., Lairson, D., Reynolds, T., Dubinsky, S., Newmark, M., Barnwell, F., Hauser, A., and Hesdorffer, D. 2011. Socioeconomic status, health care use, and outcomes: Persistence of disparities over time. *Epilepsia*, *52*(5), 957-964. doi:10.1111/j.1528-1167.2010.02968.x

Hunt, K., Weber, E., Showstack, J., Colby, D., and Callaham, M. 2006. Characteristics of Frequent Users of Emergency Departments. *Annals of Emergency Medicine, 48*(1), 1-7. doi:10.1016/j.annemergmed.2005.12.030

Krieger, N., Chen, J., Waterman, P., Soobader, M., Subramanian, S., and Carson, R. 2003. Choosing area based socioeconomic measures to monitor social inequities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project. *J Epidemiol Community Health, 57,* 186-199.

LaCalle, E., and Rabin, E. 2010. Frequent Users of Emergency Departments: The Myths, the Data, and the Policy Implications. *Annals of Emergency Medicine*, *56*(1), 42-47. doi:10.1016/j.annemergmed.2010.01.032.

Miranda, M., Ferranti, J., Strauss, B., Neelon, B., and Califf, R. 2013. Geographic Health Information Systems: A Platform to Support the 'Triple Aim'. *Health Affairs, 32*(9), 1608-1615. doi:10.1377/hithaff.2012.1199.

Patel, A., and Waters, N. 2012. Using Geographic Information Systems for Health Research. *Application of Geographic Information Systems,* 303-315. InTech. doi:10.5772/47941.

St. Sauver, J., Grossardt, B., Yawn, B., Melton, L., Pankratz, J., Brue, S., Rocca, W. 2012. Data Resource Profile: The Rochester Epidemiology Project (REP) medical records-linkage system. *Int J Epidemiol, 41*(6), 1614-24. doi:10.1093/ije/dys195.

St. Sauver, J., Grossardt, B., Yawn, B., Melton, L., Rocca, W. 2011. Use of a Medical Records Linkage System to Enumerate a Dynamic Population Over Time: The Rochester Epidemiology Project. *Am J Epidemiol*, *173*(9), 1059-1068. doi:10.1093/aje/kwq482.

St. Sauver, J., Warner, D., Yawn, B., Jacobson, D., McGree, M., Pankratz, J., … Rocca, W. 2013. Why patients visit their doctors: assessing the most prevalent conditions in a defined American population. *Mayo Clin Proc, 88*(1), 56-67. doi:10.1016/j.mayocp.2012.08.020.

Standard Hierarchy of Census Geographic Entities [Online image]. Retrieved October 12, 2014 from https://www.census.gov/geo/reference/pdfs/geodiagram.pdf.