David Fan, DBMI 2016

# Visualizing geographic trends in insurance claims data

Mentor: Dr. Raj Manrai and Professor Isaac Kohane

**Abstract**

Understanding geographic trends for chronic conditions may lead to useful insight for studying environmental risk factors for disease and improve the efficacy of public health efforts. CWAS is an online tool built in Shiny and R that allows users to visualize geographic trends in their data. The app plots maps of the USA at the county, state and region level colored by variables of the user's choice, outputs statistical tables and tables of normalized ICD9 rates. As a proof-of-concept, CWAS was used to plot maps of BRCA testing over three years using Aetna data. The subsequent plots demonstrated a county-wide increase in BRCA testing following Angelina Jolie's bilateral mastectomy in 2013.

## Introduction

Chronic diseases have complex causes that make them difficult to study; while genetics undeniably contribute to chronic disease, exposure to environmental risk factors can also elevate one's risk of developing a chronic disease.[1] Over the past decade, many studies have focused on the genetic component of chronic diseases.[2] GWAS – developed nearly ten years ago – has been used extensively by researchers as a hypothesis-generating tool to identify genetic variants highly correlated with various diseases. However, not as much attention has been given to environmental risk factors and devoting more effort to study them is essential to advance the current understanding of chronic diseases.

Incorporating analysis of geographic trends into studies of environmental risk factors would lead to better informed searches for disease correlates. Geographic regions differ in their

distribution of particulates, toxics and weather. Diseases with a clear geographic stratification may owe their effect to shared environmental features of such geographic regions. Diseases such as diabetes and obesity are known to cluster in the southern portion of the USA,[3,4] and other diseases may display this geographic preference as well. Knowing what diseases display clear geographic preferences can also make public health more efficient and preventative as more resources can be targeted toward areas with higher disposition to certain diseases.

**Methods**

A web application was created as an open-source tool for visualizing geographic trends within both user data and U.S census (2000 and 2010) and Aetna insurance-claims data. We used the Shiny framework for R and the following packages for R available on CRAN: shiny, shinysky, shinytheme, RMySql, choroplethr, choroplethrMaps, grid, gridExtra, gtable, DT and ggplot2. The app has three main features: 1) it can plot maps of the USA colored by a variable of the user's choice; 2) it can display statistical tables for the binomial exact, chi-squared, two-proportion or percent difference tests; and 3) it can display ICD9 rates normalized by population. For each of these three main features, the user may elect to visualize either the census data, Aetna data or upload his/her own dataset.

Data came from two sources – geographic and demographic data from the 2000 and 2010 U.S censuses and insurance-claims data from Aetna. County-level data for total population, black and white population percentages, median income, blue and red voting percentages, and percent below poverty was collected from the U.S Census Bureau's USA Counties database. The database is organized by subject into individual Excel files, each labeled by column with reference codes such as "POP225200D"; Population of one race - percent White alone 2000 (complete count). Columns of interest were extracted and compiled into .csv files for 2000 and

2010 using the reference codes provided in Mastdata.xls. All data was available for both 2000 and 2010 with the exception of the voting data, which was available for 2000 and 2008. The 2008 voting data was used in the 2010 .csv file.

All parameters were provided as counts, so to obtain rates for black/white population, voting and poverty, these counts were normalized by total county population. A Python script converted #DIV/0! values - which corresponded to counties with no data in either year - to a numeric value of 0. This cleansed data was then staged in a local MySQL database for 2000 and 2010.

After the above county-level data was retrieved and cleaned, state and region-level data was then produced by aggregating the county-level data. Each county of the USA has a unique five-digit FIPS code and the first two digits identify each county's state. State-level data was aggregated using this fact. Region-level data was produced using the census bureau's official definitions for the nine divisions of the USA (Figure 1).

BRCA testing counts across the USA were collected from the Aetna dataset for 2013, 2014 and 2015. As a proof-of-concept for the app, this data was fed in to study the geographic distribution of the "Jolie Effect" – the observed increase in BRCA testing following Angelina Jolie's bilateral mastectomy in 2013.

**Results**

The app homepage may be seen in Figure 2. The top red navigation bar allows the user to switch between features of the app while the area below displays options and plots. A typical user workflow looks like the one described in Figure 3. The user always starts by selecting a dataset (Census or Aetna) or uploading his or her own data. Next, the user can choose whether to plot choropleth maps, view table statistics or ICD9 rates. For each of these three main features, the user may elect to view their data at the county, state or region level. If the user chooses the

census data, the variables available to plot are the ones described above: total population, black/white population percent, median income, blue/red voting percent and percent below poverty – at the geographic level specified. If the user chooses to upload his/her own data, the app takes every column except the first to be the variables.

Figure 4 demonstrates the app's county-level plotting capabilities. Shown is a plot of the USA by median income in 2000 at the county level colored in blue. Areas around California and the Northeast are colored darker, indicative of higher relative wealth in those areas. Figure 5 demonstrates the app's state-level plotting capabilities. Shown is a plot of the USA by percent poverty at the state level colored in green. States in the south are darker, indicating higher levels of poverty relative to the rest of the country. Figure 6 demonstrates the app's region-level plotting capabilities. Shown is 1 by 2 plot of the USA by percent blue voting on the left, and percent red voting on the right, at the region level. States like California are highly liberal while states like Texas are highly conservative. Figure 7 demonstrates the app's ability to plot difference between two variables. Shown is a plot of the USA by percent difference in population between 2000 and 2010 colored on a red to green gradient, where red represents a decrease in population and green represents an increase in population. The plot demonstrates that most areas of the USA experienced an increase in population from 2000 to 2010. Figure 8 is a 3x3 plot generated using someone else's data at the county level.

Plots using the Aetna data were also generated. Figure 9 is a 3x3 plot of total BRCA tests done at the county, state and region level through 2013 – 2015. From top to bottom, the rows are 2013, 2014 and 2015. From left to right, the columns are county, state and region-level. Going down from row to row, the maps get darker for all geographic areas, regardless of geographic specificity. Two plots by percent difference from 2013 to 2014 also display this increase. In

Figures 10 and 11, green represents an increase while red represents a decrease in BRCA testing. All states but five in Figure 10 are green. When states with less than 100 total BRCA tests were filtered out in Figure 11, all states were green – the same states that were red in Figure 10 were excluded by the filter criterion in Figure 11. Figure 12 is a 1 x 3 plot of the log Jolie Effect (BRCA tests done in 2014 divided by BRCA tests done in 2013) at the county, state and region level. For all three plots in Figure 12, the log Jolie Effect curve asymptotically approaches 0.45.

**Discussion**

Figure 8 demonstrates the versatility of our app as a data-visualization tool since quality figures were successfully generated on-demand from an entirely new dataset. In the future, other people will be able to use the app in a similar manner to visualize their data. The increase in darkness across the USA from row to row in Figure 9 demonstrates the nationwide scope of the Jolie effect; regardless of whether the data is plotted at the county or region level, or where a given area is located, an increase in BRCA testing is seen from 2013 to 2014. The fact that all states were colored green after states with fewer than 100 BRCA tests were filtered out demonstrates the effect of small sample on effect size – every state with non-trivial amounts of BRCA testing demonstrated an increase in BRCA testing from 2013 to 2014. The variance plots expand on this theme. While variance is high in the county log Jolie plot, variance decreases to a minimum in the region-level log Jolie plot. The stability of the Jolie effect at lower levels of geographic specificity demonstrates the variability of perceived effects at smaller regions.

**Future Work**

While the app performs well, it requires user-uploaded data to follow a strict format and is configured currently to plot maps of the same geographic specificity at one time. These are features that can be changed to improve user experience. In addition to publishing the Shiny app

as a web app, we plan to make a R package on CRAN. The app will be applied to study more comorbid conditions in the Aetna dataset such as diabetes.

## Conclusion

CWAS is a versatile visualization tool that can be used to generate hypotheses from geographic trends. Successful application of CWAS to Aetna data confirmed the Jolie effect on BRCA testing and is but one example of how the app can be used. In the future, users are likely to use CWAS to complement their studies as a hypothesis-generating tool.

## Acknowledgements

I owe many thanks to Dr. Manrai and Kohane for taking me into the lab and allowing me to work on this highly exciting project, as well as Susanne Churchill, Barbara Mawn, Jean Fan and Dominique Altarejos for taking a chance on me by giving me my first experience in bioinformatics this summer. Finally, thanks to all the HST summer interns who shared many wonderful memories with me.

## References

1. Schwartz, D., & Collins, F. (2007). MEDICINE: Environmental Biology and Human Disease. Science, 316(5825), 695-696. doi:10.1126/science.1141331
2. Patel, C. J., Bhattacharya, J., & Butte, A. J. (2010). An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. PLoS ONE, 5(5). doi:10.1371/journal.pone.0010746
3. Barker, L. E., Kirtland, K. A., Gregg, E. W., Geiss, L. S., & Thompson, T. J. (2011). Geographic Distribution of Diagnosed Diabetes in the U.S. American Journal of Preventive Medicine, 40(4), 434-439. doi:10.1016/j.amepre.2010.12.019
4. Slack, T., Myers, C. A., Martin, C. K., & Heymsfield, S. B. (2014). The geographic concentration of us adult obesity prevalence and associated social, economic, and environmental factors. Obesity, 22(3), 868-874. doi:10.1002/oby.20502

| Region | States |
|---|---|
| New England | Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont |
| Mid-Atlantic | New Jersey, New York, and Pennsylvania |
| East North Central | Illinois, Indiana, Michigan, Ohio, and Wisconsin |
| West North Central | Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota |
| South Atlantic | Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, Washington D.C., and West Virginia |
| East South Central | Alabama, Kentucky, Mississippi, and Tennessee |
| West South Central | Arkansas, Louisiana, Oklahoma, and Texas |
| Mountain | Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming |
| Pacific | Alaska, California, Hawaii, Oregon, and Washington |

Figure 1: A table of the U.S Census Bureau's official nine divisions of the USA, used to generate the region-level data for our analysis.

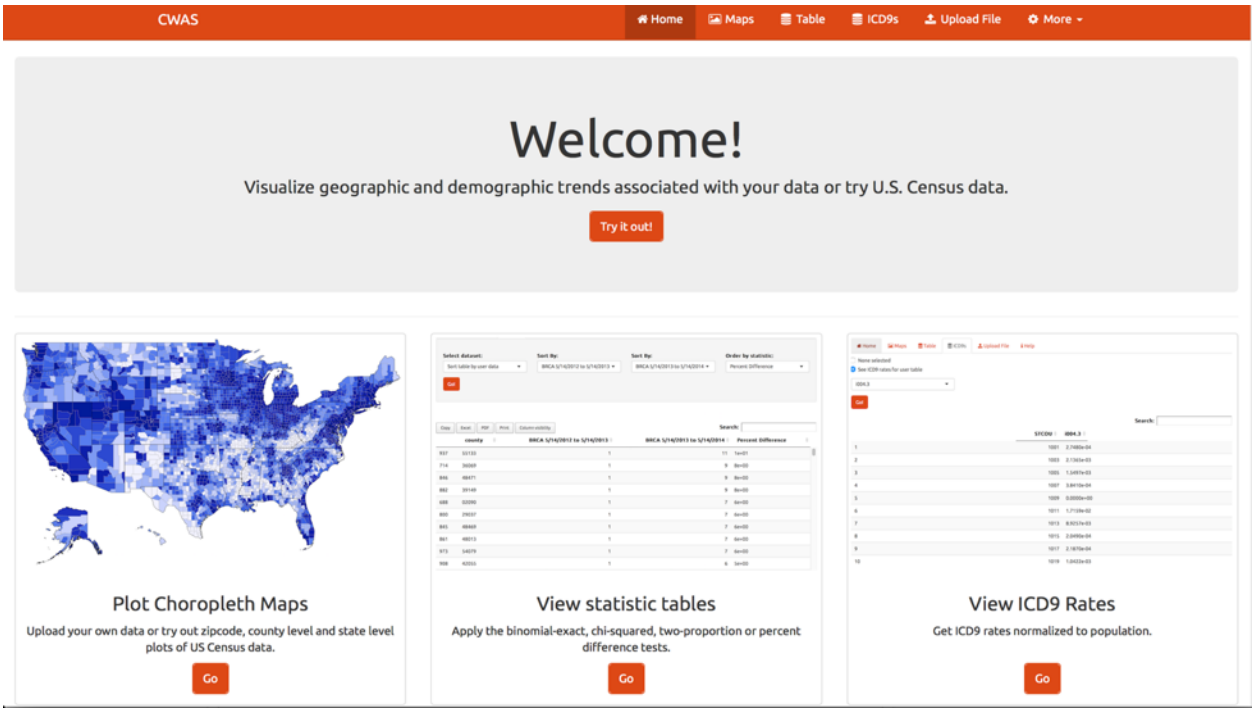


Figure 2: Homepage of the Shiny app; the red navbar allows the user to navigate between features of the app while the gray area displays the app's content.

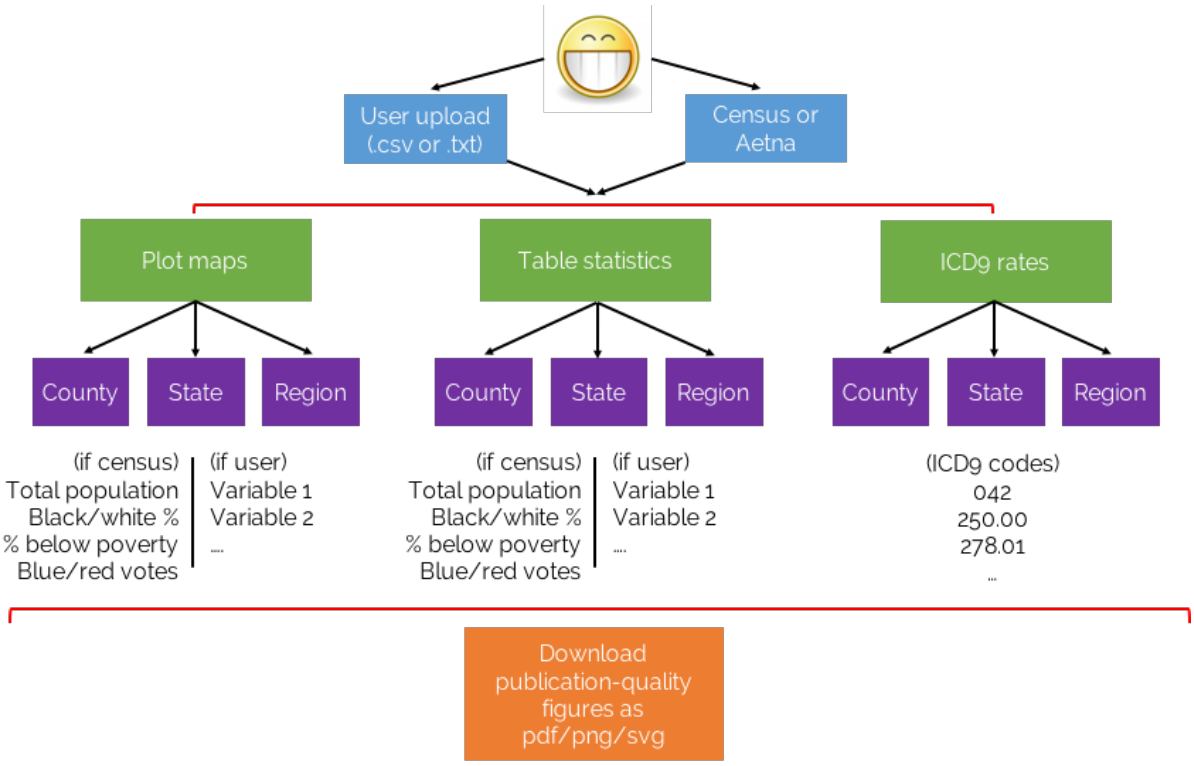


Figure 3: Demonstration of typical user workflow while using the app.

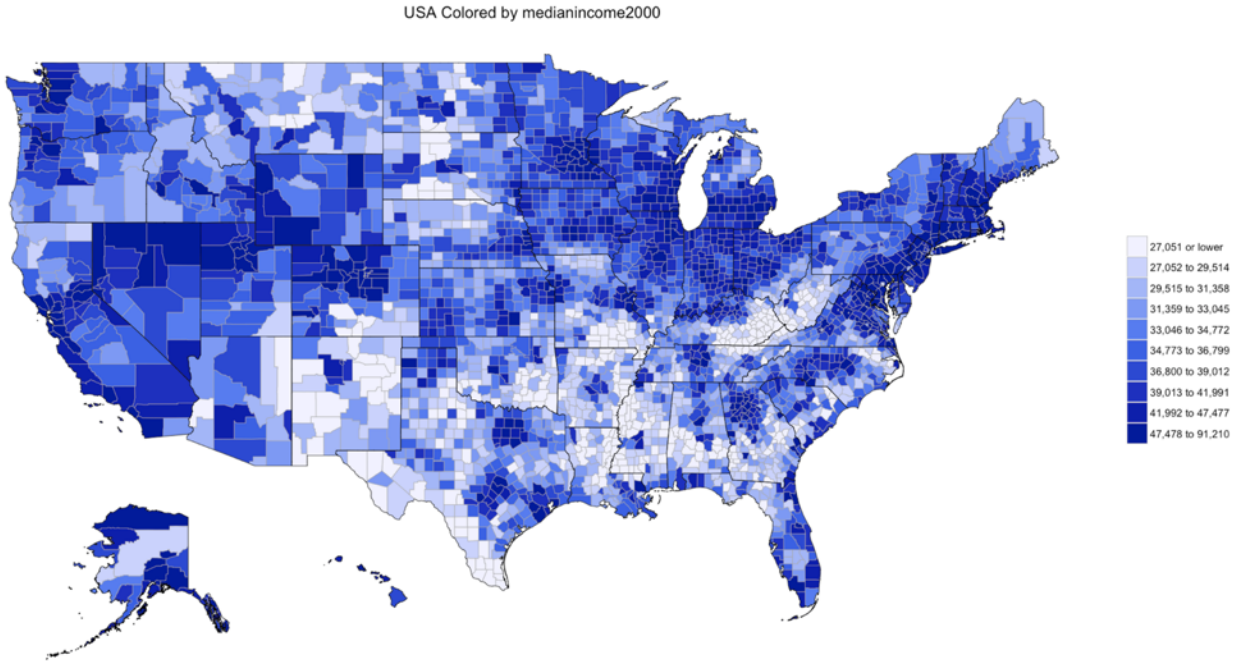


Figure 4: Plot of the USA at the county-level by median income in 2000.

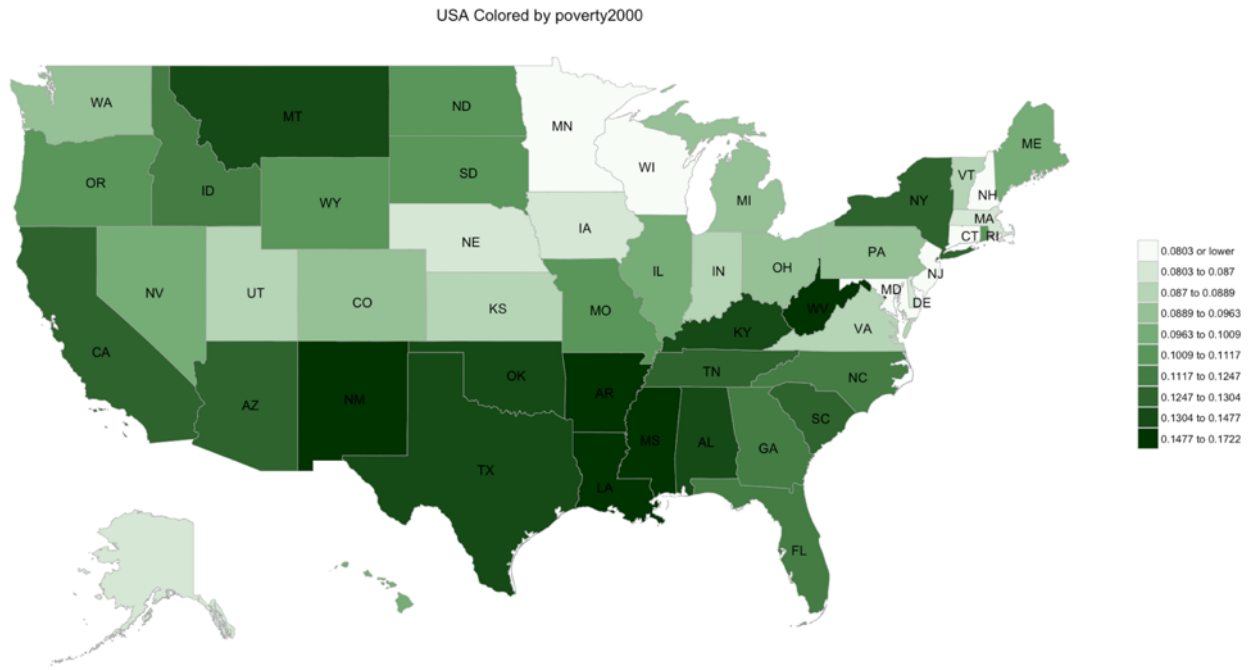


Figure 5: Plot of the USA at the state-level by percent poverty in 2000.

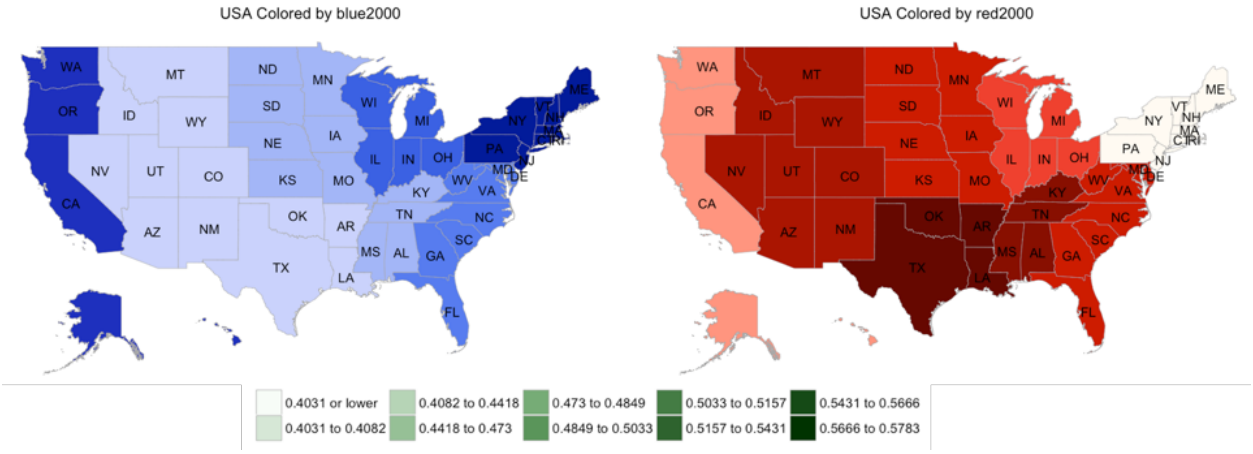


Figure 6: Side-by-side plot of the USA at the region-level by percent blue and red voting in 2000.

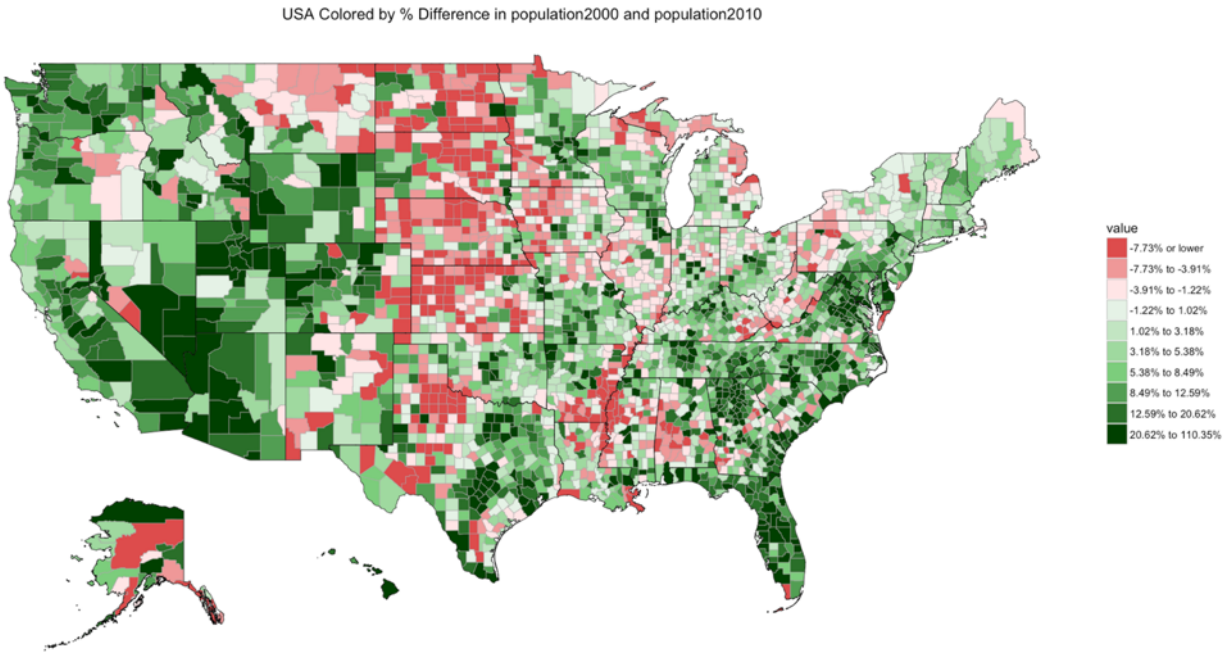


Figure 7: Plot of the USA by percent difference in population between 2000 and 2010.

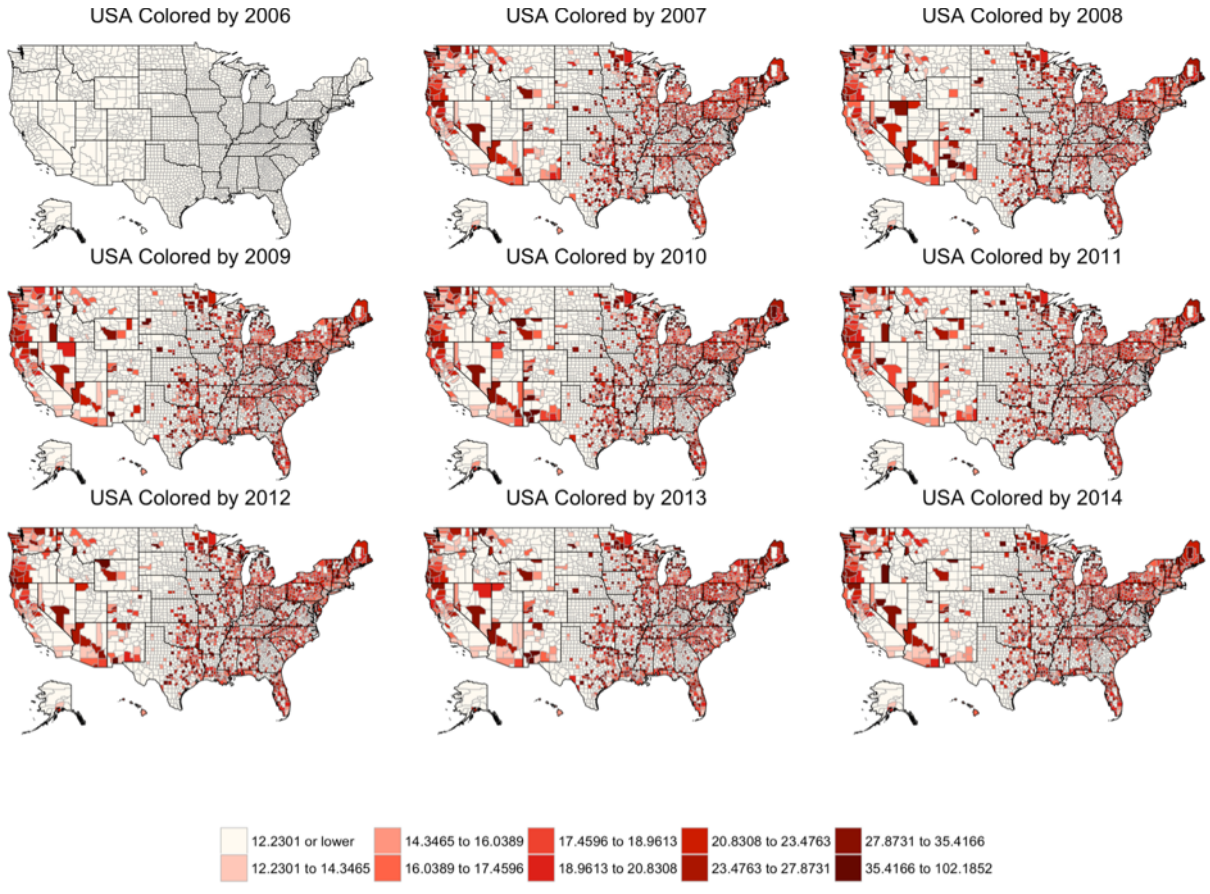


Figure 8: 3 by 3 plots of the USA at the county level using another summer intern's data from the Patel Lab.

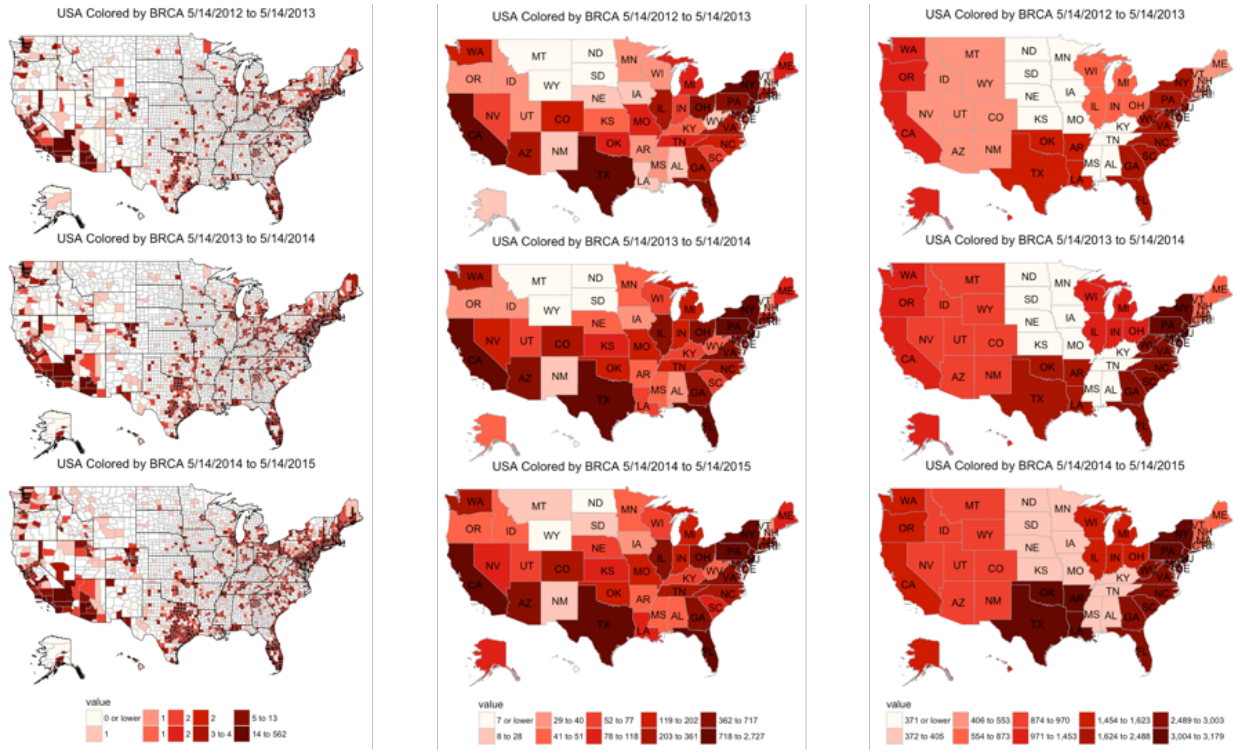


Figure 9: 3 by 3 plot of the USA at three levels of geographic specificity. From left-to-right, the columns are county-level, state-level followed by region-level. From top-to-bottom, the rows are 2013, 2014 and 2015.

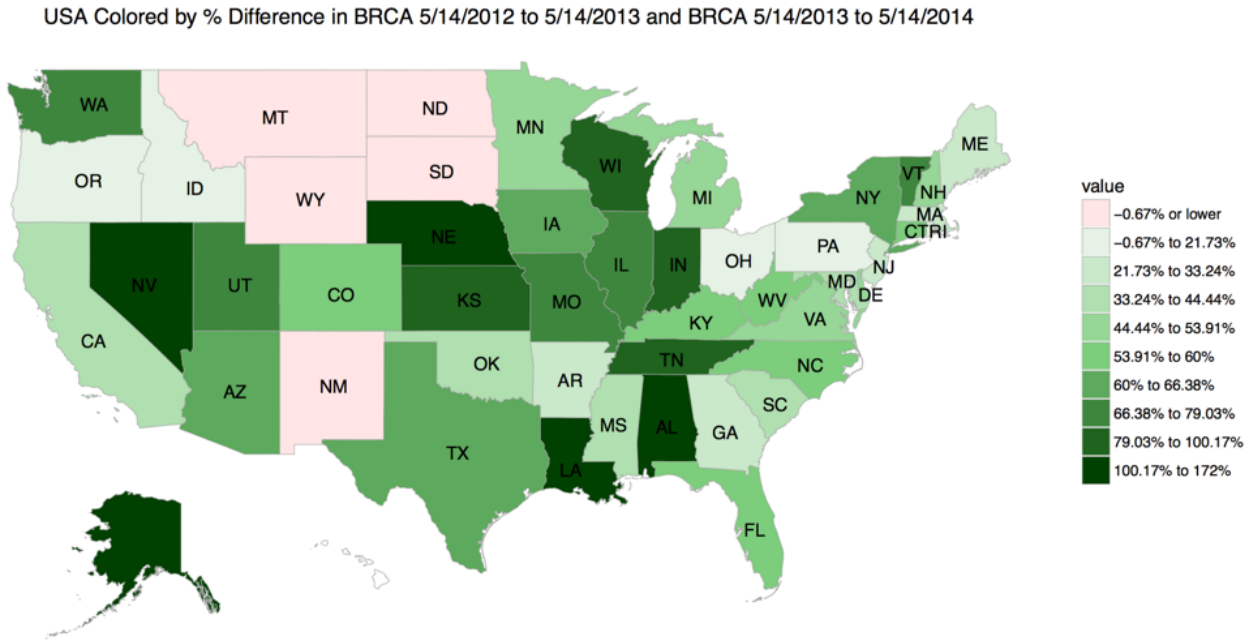


Figure 10: Plot of percent difference in BRCA testing between 2013 and 2014 at the state level.

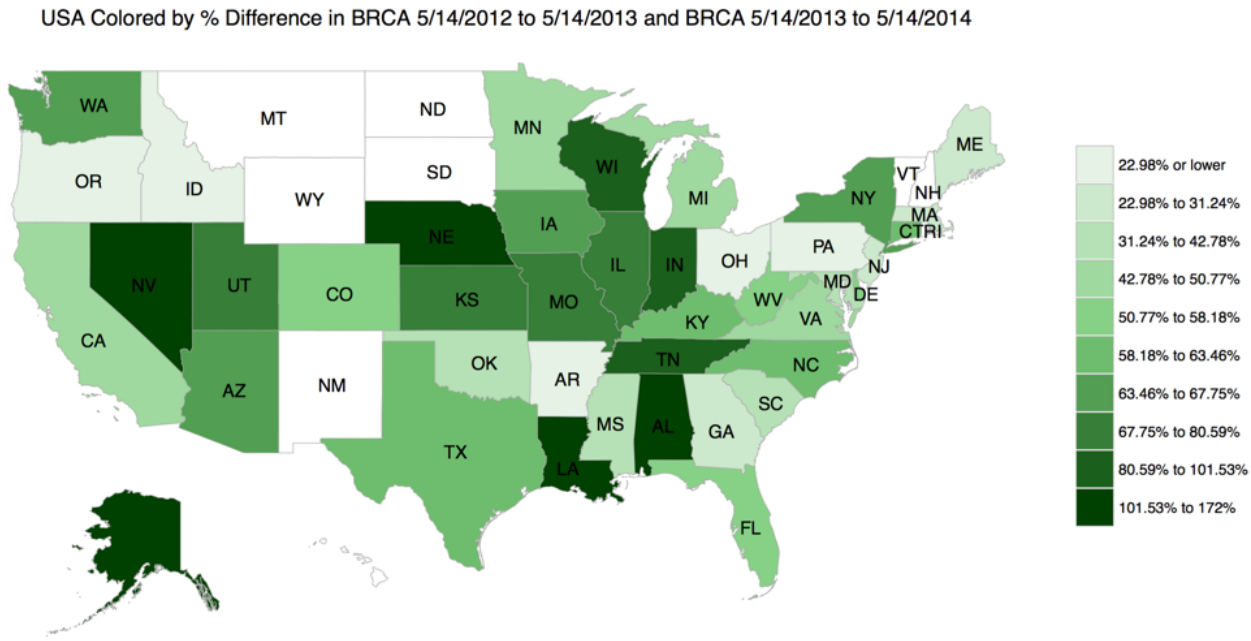


Figure 11: Same as Figure 10 except filtered to exclude states with fewer than 100 total BRCA tests performed between 2013-2015. Notice that the states excluded are exactly the ones colored red in Figure 10.

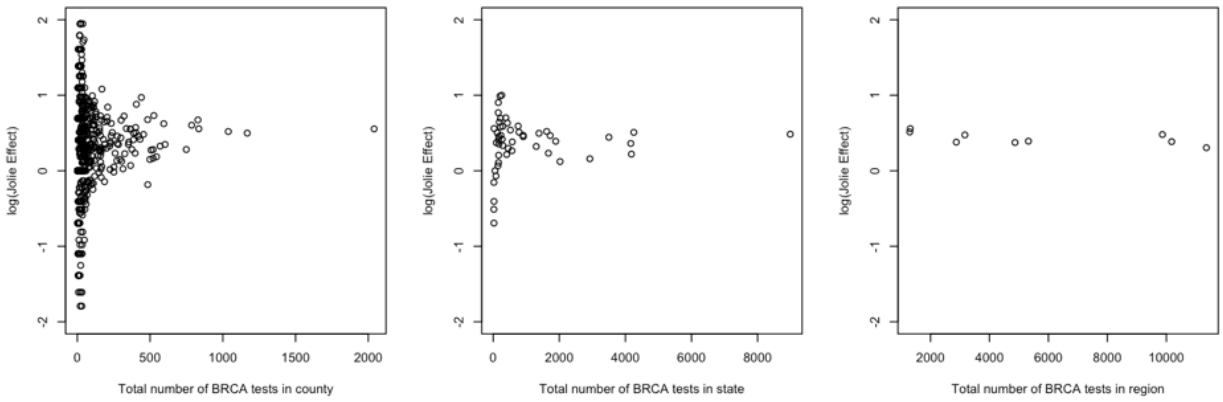


Figure 12: Plot of log(BRCA tests in 2014 / BRCA tests in 2013) at the county-level, state-level and region-level (from left to right). The region-level plot has low variance compared to the county-level and demonstrates a stable Jolie effect of 1.57 (e^0.45).