Health Access, Behaviors, and Health Outcomes Across United States Cities

Maya Ghanem and Isabelle Xiong

11/16/2021

Background and Significance

America, unlike other developed countries, does not have a universal healthcare program, meaning that not all individuals are granted free health insurance by the government. In America, Health insurance is purchased in the private marketplace or provided by the government only to certain groups, such as pregnant, low income, elderly, families with children.

The members of this project have personal experiences with United States health insurance. Having lived in Canada, where all citizens can apply for a public health insurance for free which grants them access to quality healthcare, Isabelle has never had to think of what it would be like without access to healthcare. In doing this research, Isabelle wanted to understand the effects of a lack of health insurance in the US, by finding the correlation between not having health insurance and the likelihood of contracting different diseases in 500 largest US cities. Maya worked at Westminster Free Clinic for four years, which offered healthcare services to the uninsured population in Ventura County, California. They witnessed how the obstacles our patients faced in accessing healthcare impacted their behaviors and healthcare outcomes.

Data Collection and Variables

The 500 cities dataset is provided by the Center for Disease Control and Prevention (CDC), Division of Population Health, Epidemiology and Surveillance Branch. Data from this dataset is sourced from 28,000 census tracts in "Census Bureau 2010 census population data, Behavioral Risk Factor Surveillance System (BRFSS) data (2017, 2016), and American Community Survey (ACS) 2013-2017, 2012-2016 estimates". All data was collected through the form of surveys. The meta dataset features data for a total of 5 "unhealthy behaviors", 13 "health outcomes' ' and 9 "preventive services" related to 27 types of chronic disease in 500 largest cities in the US. Within the datasets for diabetes, coronary heart disease, mental health outcomes and health insurance, variables include city, state, state abbreviation, year, datasource, TractFIPS, CityFIPS, Geographic Level, data value, low confidence unit, high confidence unit, coordinate of geographical location of city.

The original dataset includes about 29,000 observations, with multiple measurements for each city based on the data source (census, BRFSS, ACS). We restricted our dataset to only include variables on percent of city with lack of insurance, visiting the doctor, taking high blood pressure medications, smoking, reporting binge drinking, not having physical activity, with heart disease, with diabetes, and with kidney disease. We took the mean of all the age-adjusted prevalence measurements within a city and filtered our dataset to these means. As a result, there was only one measurement per city. For ANOVA testing, we filtered our dataset to only include states with measurements for at least 10 cities and created logarithmic versions of all variables.

Research Questions

1) Do cities with a greater lack of healthcare access have poorer health behaviors and/or physical health outcomes?
2) Does healthcare access, health behaviors, and/or physical health outcomes vary by state?

Variables of Interest

1) Healthcare Access for Adults (18+): Percent of City Population that Lacks Insurance, Percent of City Population with visits to doctor for routine checkup within the past year, Percent of City Population who have high blood pressure and are taking medicine for high blood pressure control.
2) Geographic Distribution by State
3) Behavior for Adults (18+): Percent of city population currently smoking, percent of city population currently reporting binge drinking habits, percent of city population reporting No leisure-time physical activity
4) Health Outcomes for Adults (18+): Percent of city population with coronary heart disease, percent of population diagnosed with diabetes, percent of city population with kidney disease

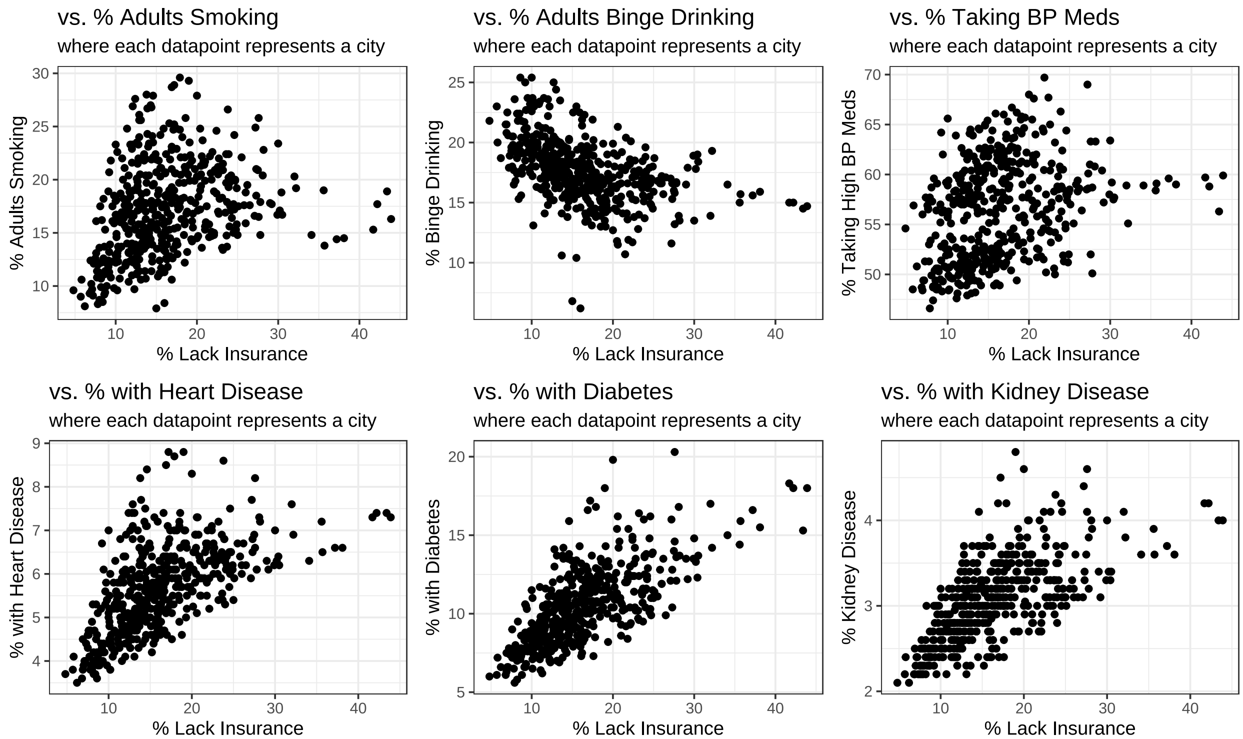
For Research Question 1, Healthcare Access Variables (1) are the explanatory variables, whereas Behavior (3) and Health Outcomes (4) are the response variables. For Research Question 2, Geographic Distribution by State (2) is the explanatory variable, and all health indicators (1, 3, 4) are the response variables.

Exploratory Data Analysis

Research Question 1

The following visuals are scatter plots between the insurance (the main explanatory variable we want to focus on) and a behavior or health outcome variable.

Figure 1: % Lacking Insurance vs. Behavior and Health Outcome Variables

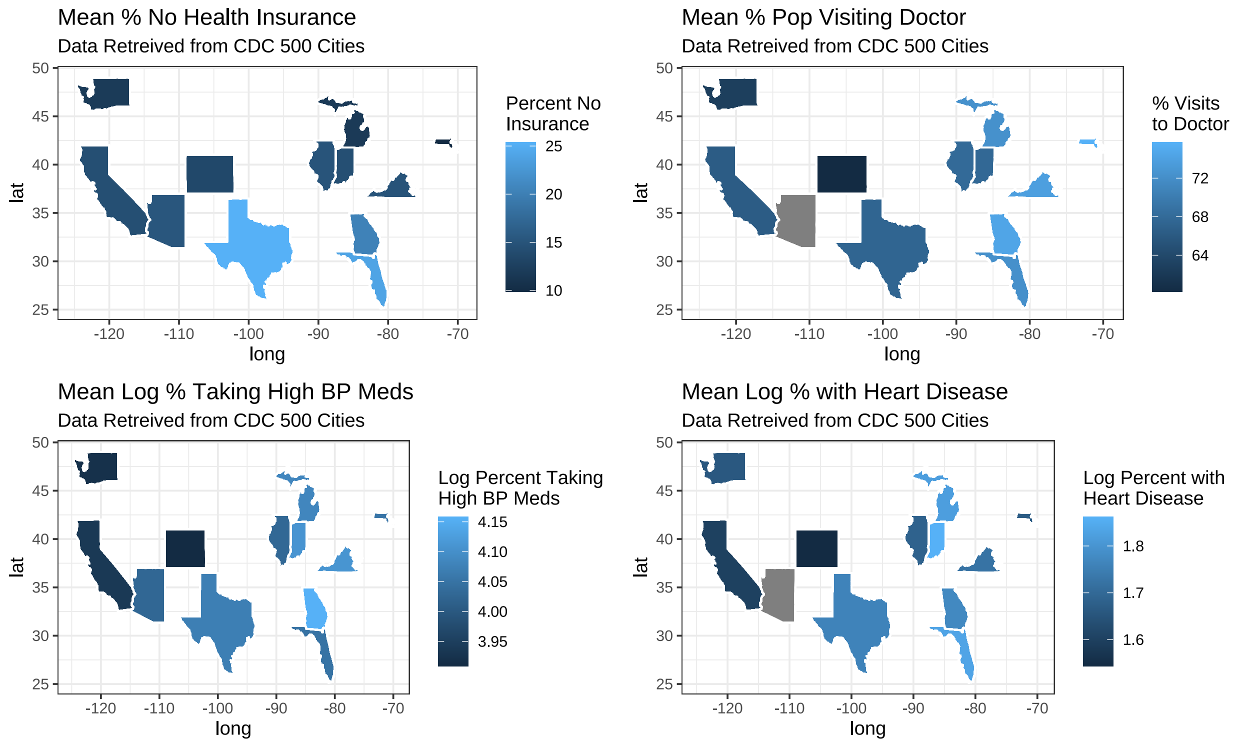


The prevalence of negative physical health outcomes have a greater positive correlation with % of population lacking health insurance than the prevalence of negative behaviors. We note that there seems to be a negative correlation between percent of city reporting binge drinking and percent of city lacking health insurance.

Research Question 2

The maps below visualize the distribution healthcare access, behavior, and outcome variables by state. Only states that have at least ten observations are included, and only four variables that fit the assumptions for ANOVA testing.

Figure 2: State Distribution of Healthcare Access, Behavior, and Health Outcome Variables



Based on the maps, it seems that Texas has the greatest percentage of people uninsured, Georgia has the greatest percentage of people taking high blood pressure medications and the highest percent of people visiting the doctor, and Illinois has the highest percentage of people with heart disease.

Analytic Methods

Research Question 1

We modeled our data with a linear regression to determine if there were significant correlations between healthcare access variables and health behaviors and physical health outcomes. We did not add any interactions to our linear models because all the explanatory variables are numerical, and the course does not encompass an analysis of interactions between two numerical variables. To determine whether a linear regression model would be appropriate, we made residual plots for each health outcome to check for any patterns in the residuals. For smoking, binge drinking, physical activity, and diabetes, there didn't seem to be a significant pattern in the residual plot, so a linear regression model was appropriate. However, for coronary heart disease and kidney disease, there was a pattern in the residual plot where the higher the predicted percentage of adults with diabetes, the larger the magnitude of the residual, such that a linear regression model would not be appropriate.

Research Question 2

We conducted ANOVA testing to determine if there was variance within and between states for the healthcare access, behavior, and health outcome variables. We determined that state distributions of percent lack of insurance, percent visiting the doctor, log percent with heart disease, and log percent taking high BP medications had normal distributions within groups and homoscedastic variance, therefore fitting the assumptions for ANOVA testing. It is highly unlikely that the samples groups within an ANOVA test are independent; for example, certain states may share similar political contexts and could be more likely to share simliar health outcomes. We employed a Bonferroni correction with a subsequent significance level of p

= 0.000641025641.

Results

Research Question 1

1) Holding all other variables constant when evaluating each explanatory variable, we expect a 0.0523 percentage increase in adults smoking in a city for every percent increase in a city lacking access to health insurance, a 0.0966 percentage decrease in adults smoking in a city for each percent decrease of people in a city visiting doctor for routine checkup within the past year, and a 0.674 percentage increase in adults in a city smoking for each percent increase in adults in city with high blood pressure had access to high blood pressure medication. The coefficients for all three variables are statistically significant (p-value < 0.05), meaning there is less than a 5% chance such a coefficient or more extreme would be found in the data if percentage of adults smoking and the healthcare access variables were not associated.

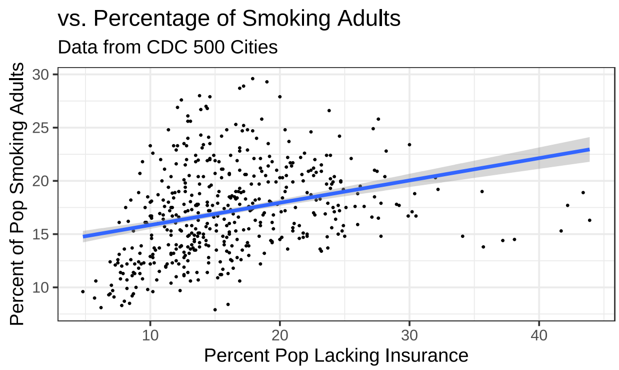
2) Holding all other variables constant when evaluating each explanatory variable, we expect a 0.162 percentage decrease of adults in a city reporting binge drinking for every percent increase of adults in a city lacking access to health insurance, a 0.0565 percentage increase in adults in a city reporting binge drinking for each percentage increase of adults in a city visiting doctor for routine checkup within the past year, and a 0.137 percentage decrease in adults in a city reporting binge drinking for each percent increase in adults in a city with high blood pressure had access to high blood pressure medication. The coefficient for all three variables are statistically significant (p-value < 0.05).

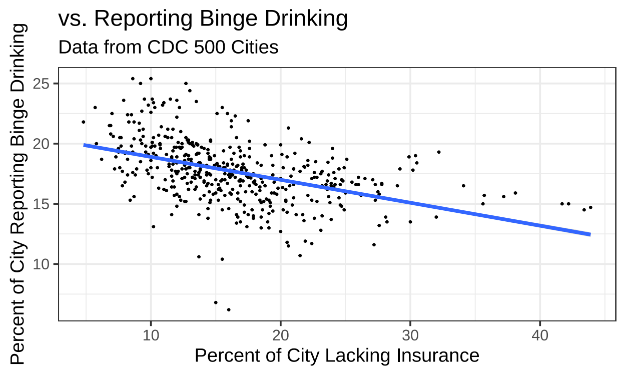
3) Holding all other variables constant when evaluating each explanatory variable, we expect a 0.0669 percentage increase of adults in a city reporting no physical activity for each percent increase of adults in a city lacking health insurance, a 0.0113 percentage decrease in of adults in a city reporting no physical activity for each percent increase of adults in a visiting doctor for routine checkup within the past year, and a 0.1220 percentage increase of adults in a city reporting no physical activity for each percent increase of adults in a city with high blood pressure had access to high blood pressure medication. The coefficient for all three variables are statistically significant (p-value < 0.05).

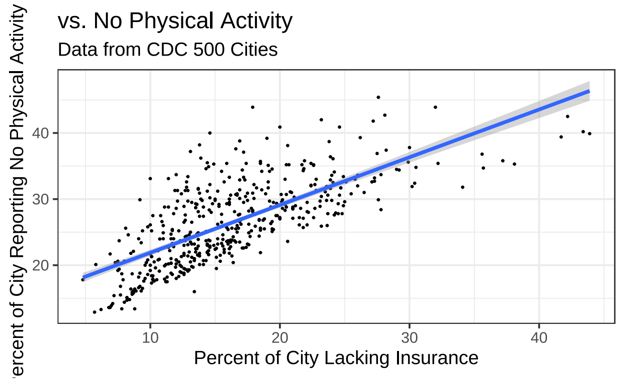
4) Holding all other variables constant when evaluating each explanatory variable, we expect a 0.239 percentage increase in population diagnosed with diabetes for each percent increase of adults in a city lacking health insurance, a 0.065 percentage decrease in population diagnosed with diabetes for each percent increase of adults in a visiting doctor for routine checkup within the past year, and a 0.171 percentage increase in population diagnosed with diabetes for each percent increase of adults in a city with high blood pressure had access to high blood pressure medication. The coefficient for all three variables are statistically significant (p-value < 0.05).

Figure 3: Percent Lacking Insurance vs. Response Variables (Titled in Quadrants)

Linear Regression Considers Visits to Doctor and High BP Meds as Explanatory Variables







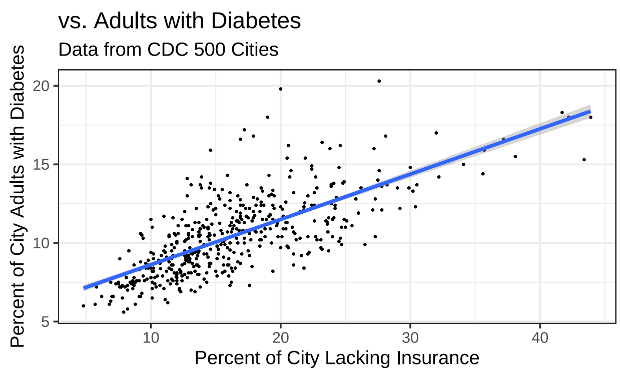


Table 1: Intercepts and Coefficients for Linear Regressions

Coefficients	Smoking	Binge Drinking	Physical Activity	Diabetes
Intercept	-15.0000	24.2000	-28.1000	-7.570
Insurance	0.0523	-0.1620	0.5330	0.239
Visits to Doctor	-0.0966	0.0565	0.0625	0.065
High BP Meds	0.6740	-0.1370	0.7380	0.171

Table 2: P-Values of Intercepts and Coefficients for Linear Regressions

Significance Value	Smoking	Binge Drinking	Physical Activity	Diabetes
Intercept P-Value	0.0000000	0.0000000	0.0000000	0.0000000
Insurance P-Value	0.0279000	0.0000000	0.0000000	0.0000000
Visits to Doctor P-Value	0.0308000	0.0945000	0.0995000	0.0021300
High BP Meds P-Value	0.0000000	0.0000439	0.0000000	0.0000000
Adjusted R-Squared Value	0.5150724	0.2367489	0.8369087	0.6797326

Research Question 2

ANOVA tests provided evidence that there are significant variances between and within states for percent lack of insurance, percent visiting the doctor, log percent with heart disease, and log percent taking high BP medications. The step down tests demonstrated that there are 19 significant pairs for percent lack of insurance, 47 significant pairs for percent visiting doctor, 39 significant pairs for log percent taking high BP meds, and 7 significant pairs for log percent with heart disease. Florida and Texas had the most significant pairs (9 each) for the insurance step down tests, Colorado and Washington had the most significant pairs (10 each) for the visits to doctor step down tests, California, Colorado, Georgia, and Washington (9 each) had the most significant pairs for the log medicine high BP step down tests, and Colorado had the most significant pairs (3) for the log % heart disease step down tests. The only step-down test that had states with no significant pairs was for log percent heart disease.

Table 3: ANOVA Summary Table Including F-Statistics and Significant Pairs

...1	Insurance	Visits to Doctor	log Medications	log Heart Disease
F-Statistic (Overall)	25.2	90.03	113.3	10.3
Overall Sig Pairs	19.0	47.00	39.0	7.0
Arizona Sig Pairs	2.0	7.00	6.0	0.0
California Sig Pairs	2.0	7.00	9.0	4.0
Colorado Sig Pairs	2.0	10.00	9.0	3.0
Florida Sig Pairs	9.0	8.00	6.0	2.0
Georgia Sig Pairs	1.0	7.00	9.0	0.0
Illinois Sig Pairs	2.0	8.00	6.0	0.0
Indiana Sig Pairs	2.0	6.00	6.0	2.0
Massachusetts Sig Pairs	3.0	9.00	4.0	0.0
Michigan Sig Pairs	2.0	8.00	3.0	2.0
Texas Sig Pairs	9.0	7.00	4.0	1.0
Virginia Sig Pairs	2.0	7.00	6.0	0.0
Washington Sig Pairs	2.0	10.00	9.0	0.0

Discussion

Research Question 1

The linear regression models demonstrate that there is generally a positive correlation between health insurance and negative behaviors and physical health outcomes. This result matches the findings of Zuckerman's article that a lack of health insurance impacts other health outcomes. The only exception is that one linear regression suggested decreased binge drinking as healthcare access decreases. This is counter intuitive to the idea that less access to health care results in worse behavior outcomes.

The adjusted R-square values for Smoking, Diabetes and Physical activity are above 0.5, meaning that more than 50% of the variability in the outcome data can be explained by the linear model. The adjusted R-square value for Binge Drinking is 0.2367, suggesting that only around 24% of the variability in the outcome data can be explained by the linear model.

It is highly unlikely that the explanatory variables are independent; for example, the fact that a person who has insurance is more likely to visit the doctor and take medications demonstrates that these groups are not independent. A way to remedy this would be to include interactions between correlated explanatory variables, but interpreting interactions between two numerical variables is beyond the scope of this course.

Further research could look into causes for the correlations between the variables in the linear regressions. In particular, further research should investigate why there is a negative correlation between percent of city lacking health care access and percent of city binge drinking. A different type of regression model would likely be more appropriate and have higher adjusted R-square values closer to 1. Further research should look into different types of regression models for the data, but this is beyond the scope of the course.

Research Question 2

The ANOVA testing demonstrated that there is variance within and across states for lack of insurance, visits to doctor, heart disease, and taking High BP medication rates. Based on their higher F-squared values and greater number of significant pairs, visits to doctor and High BP Medication rates seemed to have more variance than lack of insurance and heart disease.

These results are significant in supporting Zuckerman's article on how there are differences in insurance coverage across states; however, our exploratory data refutes the claim that California, Texas, and Colorado having disproportionately high uninsurance rates, instead indicating that Texas has teh highest uninsurance rates (Zuckerman 1999, 8). These variances for health insurance correspond with state based variances for doctors' visits, heart disease prevalence, and people taking high blood pressure medications.

A Bonferroni correction is implemented on the ANOVA testing because it is unlikely that the state distributions are independent from each other, but the research is limited in identifying what exactly causes a lack of independence. Identifying this cause could allow one to correct the lack of independence more accurately.

Further research should investigate the socioeconomic and policy factors that contribute to differences between and within states. Are the differences due to differing statewide healthcare policies or the result of inherent economic inequality between states?

References

https://chronicdata.cdc.gov/500-Cities-Places/500-Cities-Local-Data-for-Better-Health-2019-relea/6vp6-wxuq

https://www.urban.org/sites/default/files/publication/66251/309311-Snapshots-of-America-s-Families.PDF

https://www.annualreviews.org/doi/abs/10.1146/annurev.publhealth.28.021406.144042