BAIS:3250 Final Report

Education's Impact on Cardiovascular Disease Development

Cassady Jackson and Baylie Schnieder

GitHub Repository: https://github.com/bayliemschnieder/CVD-VERSUS-EDUCATION

Introduction

This report aims to better understand the health impacts of lower quality education in the United States. According to the Institute for Health Metrics and Evaluation[1], "18 years of education reduces the risk of death by 34% and is comparable to eating a healthy diet." This same study found that not going to school is as bad for your health as drinking five or more alcoholic drinks per day or smoking ten cigarettes a day for ten years. Based on this information, we would like to explore the rates of chronic illness in each state using their public education rankings. These health impacts could be due to the riskier lifestyles that young teens who are not in school are more likely to lead. We would like to explore the idea that schools ranked higher in education may have lower rates of chronic illnesses.

In our initial research, we realized that the scope of the dataset was too large to provide meaningful insights into the realm of chronic illness. The data was then filtered to include only cardiovascular disease because, as of 2022, according to the CDC, heart disease was responsible for one in every five deaths in the United States[2]. Throughout this report the findings of the analysis will be detailed, as to whether there is an impact on cardiovascular disease rates per every 100,000 residents based on state education rankings.

Data

[1] https://www.healthdata.org/news-events/newsroom/news-releases/learning-life-higher-level-education-lower-risk-dying#:~:text=18%20years%20of%20education%20reduces,per%20day%20for%2010%20years.

[2] https://www.cdc.gov/heart-disease/about/?CDC_AAref_Val=https://www.cdc.gov/heartdisease/about.htm

Our first data source is from data.gov[3] and includes information on Chronic Health Indicators from all 50 states collected from 2018 to 2022. This source offers us CSV, RDF, JSON and XML files to work with, so we will be able to choose the one that works best with our scraped data. This data includes information on the year that it was collected, the state that it was collected from, what the source of the data is, what type of disease each symptom indicates, what questions the individuals were asked during collection and what their main symptom indicator is. There is also information on the answers individuals gave, what topic they were questioned on and the type of data collected. This is important because it is a compilation of data from the CDC, the Council of State and Territorial Epidemiologists and the National Association of Chronic Disease Directors. This data is meant to help states define, collect and report chronic disease information to improve medicine practice.

Our second data source is from worldpopulationreview.com[4] and includes information on public education rankings and performance in all 50 states. This is the source that we will be scraping and crawling as we embark on our data journey. We chose this source because the information provided is based on an Consumer Affairs[5] article. This Consumer Affairs article compiles information from the U.S. Census, College Board, the National Education Association and the National Center for Education Statistics, among other sources. Using all of this information they are able to make informed decisions about state education ranking, which is reported in an easy-to-read table. We were not able to scrape the Consumer Affairs article because their website does not allow crawling. World Population Review includes the same table on their website as Consumer Affairs. This table includes the state's overall education rank, higher education quality rank and school safety rank. We will use these factors to determine education levels in the school and relate them to the chronic disease indicators table.

The World Population Review, as well as Consumer Affairs, update their websites every year to reflect the most current state education rankings. In order to compare rankings from years past, we used the Wayback Machine[6] to look at the archive from the World Population Review,

[3] https://catalog.data.gov/dataset/u-s-chronic-disease-indicators

[4] https://worldpopulationreview.com/state-rankings/public-school-rankings-by-state

[5] https://www.consumeraffairs.com/movers/best-states-for-public-education.html

[6] https://web.archive.org/web/20220306220650/https://worldpopulationreview.com/state-rankings/public-school-rankings-by-state

from which we were able to pull information from 2021, to align with the last year of data collected for cardiovascular disease in the chronic health indicators file. We used this scraped data, as well as a filtered version of the Chronic Health Indicators file, including only information from 2021 and related to cardiovascular disease, to perform the analysis for this project.

These datasets were merged using an inner merge, on the “State” variable, which was included in both sets. An inner merge was chosen because of the shared state column that did not align exactly with every entry from the Chronic Health Indicators file. This merge allowed for the data from the Chronic Health Indicators to remain unchanged, while adding the education ranking information. The original dataset for Chronic Health Indicators included over three million missing values which needed to be cleaned from the dataset. For columns that we intended to use in analysis, we removed just the rows containing missing values. After that, we removed all columns containing over half missing values, “Response”, “DataValueFootnoteSymbol”, “DataValueFootnote”, “StratificationCategory2”, “StratificationCategory3”, and “ResponseID” were all dropped. Following this cleaning, analysis started on the whole file, where we discovered it would be hard to use the data to derive meaningful insights. The “Topic” column was then cleaned, to retain only values corresponding to “Cardiovascular Disease.” Following this cleaning, the “DataValueUnit” column was also cleaned, retaining only rows that corresponded to “cases per 100,000” to allow for fair comparison across states.

Data Dictionary:

Field	**Type**	**Description**
YearEnd	Numeric	End year
LocationAbbr	Text	State data collected from abbreviation
State	Text	The state name fully
DataSource	Text	Source the data is from
Topic	Text	Indicators of Disease
Question	Text	Question used in data collection
DataValue	Numeric	Number of cases

DataValueUnit	Text	Cases per 100,000
DataValueType	Text	Crude Rate or Age-Adjusted Rate
StratificationCategory1	Text	Questions asked about the participant
Stratification1	Text	Answer to the question asked to participant
LocationID	Numeric	Location identifier
TopicID	Text	Topic identifier
QuestionID	Text	Question identifier
DataValueTypeID	Text	Data value type identifier
StratificationCategoryID1	Text	Stratification category identifier
StratificationID1	Text	Stratification identifier
Overall Public School Rank 2021	Text	Overall ranking based on all other factors
Higher Ed Quality 2021	Text	Overall ranking based on universities.
School Safety Rank 2021	Text	Ranking of the safety of the schools in the state

Analysis

Answering these questions helped us to understand the impact of education on a state's overall health. They give an idea of the bigger picture of the problem we are looking to solve, which is if education level has a sizeable impact on cardiovascular disease. In earlier iterations of the project, tests were run on the entire Chronic Health Indicators dataset, which returned incredibly low accuracy ratings, -7.0562e-05 from a linear regression, and scatterplot distributions that did not have meaningful findings.

Analysis Question 1: Do state education rankings impact the rates of cardiovascular disease between states?

When curating the project, it was assumed that state education rankings would impact the rates of cardiovascular disease within the state. Our hypothesis was that state education ranking did have an impact on the rate of heart disease in a state. In testing the hypothesis, we used Welch's t-test, to compare the data values from the top 10 ranked states to the data values from the bottom 10 ranked states. This returned a t-statistic of -2.9385 and a p-value of 0.0034.

Because of how small the p-value is, there is a significant difference in data values between the top and bottom states, and we can go ahead and reject the null hypothesis. To compare the mean of all states, 125 cases per 100,000, to the mean of the top ten states, 112 cases per 100,000 and the mean of the bottom ten states, 140 cases per 100,000, there is some variance depending on the overall education ranking. The same goes for the median, which was 80 for all states, but 66 for the top states and 89 for the bottom states. This range of values helps to support the findings of the hypothesis test, that overall state education ranking does have an impact on the rate of cardiovascular disease.

To further investigate potential relationships between education rankings and rates of cardiovascular disease, we also implemented a k-nearest neighbors test. This test was used to understand the accuracy of groupings between rankings, and curate an f-1 score to determine how well groupings performed when predicting new values. This returned an f-1 score of 0.45, which is below average, meaning that overall, the groupings are not performing well when grouped based on all rankings, not just top or bottom. This means that between the top and bottom ten states there is a significant difference, but there is not as clear of a difference between all states with all rankings.

To understand the distribution among states, we chose to use a word cloud to understand what states are reporting the highest number of cases. *Figure 1*, below, is the result of the word cloud creation. The biggest states reporting based on the word cloud are Washington, California and Arizona. These are all states with the highest population, reporting cases per 100,000. Washington ranks 21st overall, 26th in higher education quality and 5th in school safety. California ranks 27th overall, 38th in higher education quality and 32nd in school safety. Arizona ranks 49th overall, 50th in higher education quality and 31st in school safety. Washington ranks in the top 25 overall, but all three states are not included in the top 10. This supports the idea from the t-test that there may be a relationship between overall state education ranking and rate of cardiovascular disease.

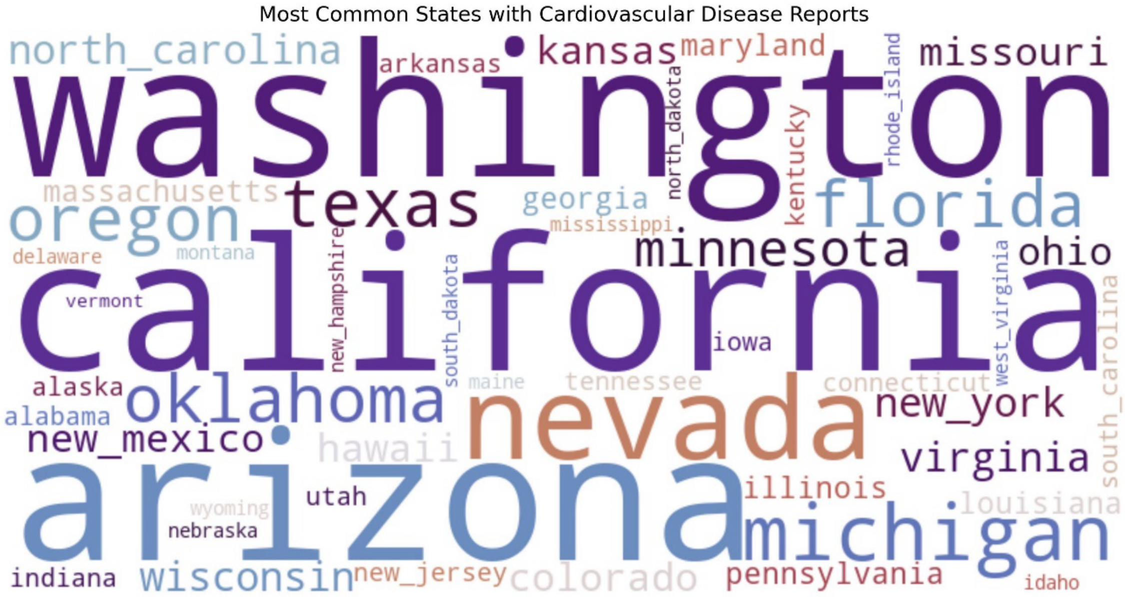

Figure 1: Word Cloud by State Appearance

Analysis Question 2: What is the distribution of chronic illness among the 10 lowest ranked states? What is the distribution of chronic illness among the 10 highest ranked states?

Due to the significant p-value difference between top 10 and bottom 10 states when comparing rate of cardiovascular disease, tests were run to understand the difference between the two. To test the existing correlation between chronic illness rates and state ranking, we divided the top 10 states and bottom 10 states into their own data frames. A Pearson correlation test was run to understand the strength and direction of a potential linear relationship between variables. The hypothesis was that there is a moderate linear correlation between overall public-school rank and the rate of cardiovascular disease per 100,000 people.

For the top ten states, we were able to find that the Pearson correlation coefficient was 0.0513, inferring that there is a weak or no linear correlation between the values. The top ten states also resulted in a p-value of 0.2538, meaning that we should reject the hypothesis that there is a relationship between state ranking and cardiovascular disease level. For the bottom ten states, we found a Pearson correlation coefficient of -0.0430, this means that there is a weak or no linear correlation. The bottom ten states resulted in a p-value of 0.3159, meaning that we should reject the hypothesis that there is a relationship between ranking and chronic illness level.

To further investigate these findings, bar charts below were used to graph the cardiovascular disease rate per 100,000 people. *Figure 2* shows the rate per 100,000 for the bottom 10 states, and *Figure 3* shows the rate per 100,000 for the top ten states. These can be used to compare the highest rate for the top ten states to the highest rate for the bottom ten states. There is some difference between the two, with the top ten states' highest rate being 175.5 for the state ranked 5th overall, and the bottom ten states' highest rate being 177.7 for the state ranked 46th overall.

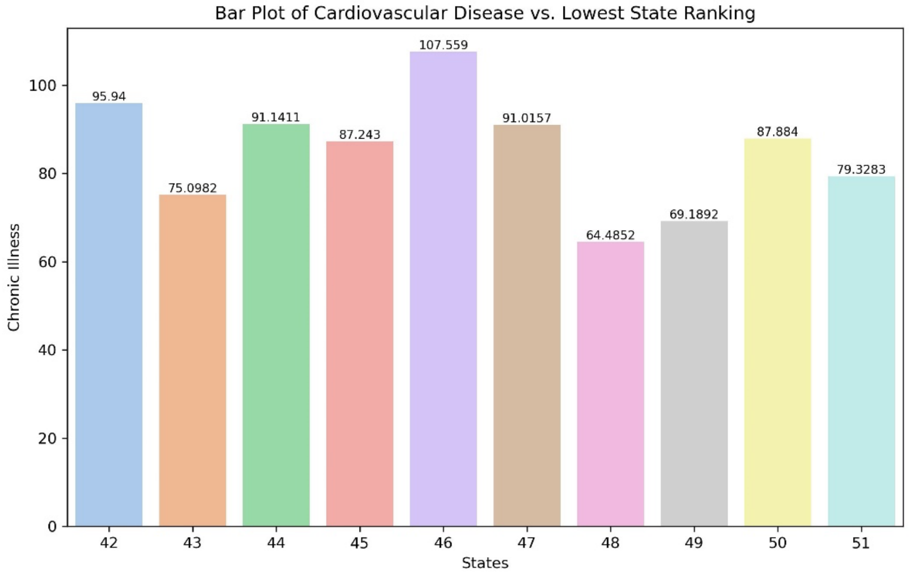

Figure 2: Bottom 10 States by Overall Education Ranking, Rate is per 100,000 people.

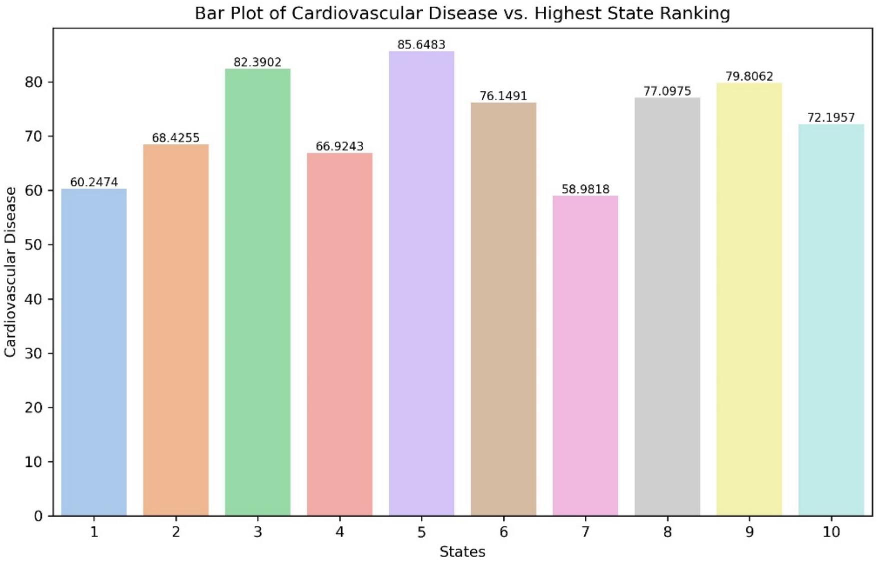

Figure 3: Top 10 States by Overall Education Ranking, Rate is per 100,000 people.

Analysis Question 3: Does education impact crude rate and age adjusted rate of cardiovascular disease development differently?

For crude rate and age adjusted rate of cardiovascular disease, we would assume that the crude rate would be higher for the states who rank lower in education. This is because we assume that age adjusted rate has more outside factors contributing to the development than crude rate. In order to test this hypothesis, we also used a Welch's t-test. This separation only compared age adjusted rate and crude rate cases, and returned a p-value of nearly zero, meaning that there is a statistically significant difference between crude rate and age adjusted rate. To further explore this, we used a logistic regression, to attempt to classify states as reporting using age adjusted rate versus crude rate. The logistic regression did not have great accuracy, with a score of 0.56, only marginally better than flipping a coin. Based on this accuracy score, we can assume that overall education score, higher education ranking and school safety ranking are not good predictors of the way that states report their cardiovascular disease data.

To better visualize the difference between crude rate reporting and age adjusted rate reporting, we created side by side scatterplots with regression lines. These served as a way to better visualize trends throughout each state, ranked based on their overall performance. As you can see in *Figure 4,* there is a little bit of a difference between the rates for the higher ranked states and lower ranked states.

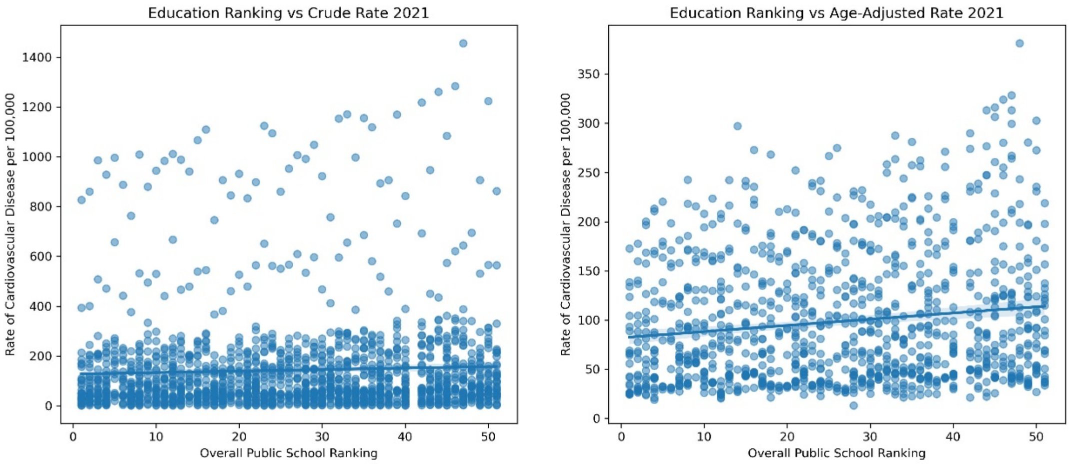

Figure 4: Cardiovascular Disease Reporting based on Overall School Ranking

Analysis Question 4: Does higher education quality impact the rate of cardiovascular disease development?

Compared to overall education quality, we would not assume that higher education quality would have an impact on the rate of cardiovascular disease development. In order to test this hypothesis, we compared the twenty-five states ranked highest in higher education to the 25 states ranked lowest in higher education's rate of cardiovascular disease development per 100,000. We used a Welch's t-test for this as well, which returned a t-statistic of -1.7304 and a p-value of 0.0837. Therefore, there is not a statistically significant difference between the rate of cardiovascular disease development between higher ranked states in higher education and lower ranked states. When looking at the descriptive statistics for the top and bottom 25 states, these findings make sense. The median values are only 7 cases apart, with their first and third quartiles being about the same as well. The bottom 25 ranked states do have slightly higher values overall, but their distributions are similar, so the lack of statistical signifiance makes sense.

To understand the distribution between the top 25 and bottom 25 states we used a color-coded scatterplot to easily compare. This scatterplot, pictured in *Figure 5,* showed that the majority of the cardiovascular diseases reported among all states were around the same rate. There were some higher outliers for the states ranked worse in higher education quality, but there is not a clear difference.

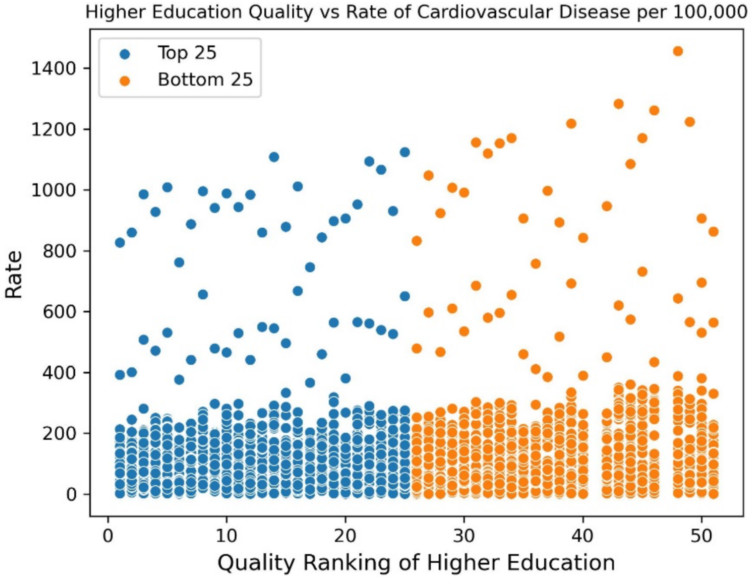

Figure 5: Higher Education Quality versus Rate of Cardiovascular Disease

Conclusion

There appears to be some relationship between cardiovascular disease development and overall education ranking between the top and bottom ten states. However, there does not appear to be a further relationship between overall education ranking and cardiovascular disease development when comparing across all states. There also does not appear to be a relationship between higher education quality and cardiovascular disease development, or crude reporting versus age adjusted when it comes to state education rankings. These findings make some sense, due to the variation among health in states which is not always impacted by education. It is not what we were anticipating finding at the beginning of researching, but with the analysis we completed it began to make sense.

If this project were to be repeated, a more effective way of collecting data should be implemented on new machine learning techniques. This data needed several cleaning measures in order to be used for testing. In future iterations, a group could collect this data themselves on the chronic illnesses so that understanding certain variables within the data wouldn't be as confusing. It would also be better to have a more recent year of data, but there has not been a recent update for 2024 or 2025. Overall, this project delivered different aspects of descriptive statistics, visualizations, hypothesis tests, machine learning techniques, and text analytics specifically word cloud. Even though results were not what we hypothesized this still showed that there could be some relationship between chronic illness, specifically cardiovascular disease, and education ranking. If run with updated data, it could have more definitive results.