COVID-19 Death and Vaccination Rates in the United States

An analysis on nationwide COVID death rates, contributing conditions, and vaccinations.

By William Donnell-Lonon, Gabriel Rivera, Colin Partridge, and Benjamin Tutka

- Date of submission

**Abstract**

The COVID-19 pandemic, became a global crisis with profound impacts. With this we prompted an investigation using CDC data. Analyzing datasets in Tableau and Python, the study faced challenges due to inconsistencies and errors in CDC data, leading to reliance on calculated totals. Four dashboards were created to explore death trends, contributing conditions, race correlations and the correlation between vaccines and death rates.

The first dashboard revealed that southeastern states exhibited higher per-capita death rates, linked to potential factors like lower socioeconomic conditions, vaccine hesitancy, and government distrust. Notably, older populations were significantly more vulnerable to COVID-19. The second dashboard delved into contributing conditions, finding respiratory diseases and circulatory conditions as major contributors, affecting different age groups differently. The third dashboard explored the correlation between vaccine rates and COVID deaths over time, indicating a relationship between awareness and vaccination. The final dashboard provides advanced user intractability, showing how variables such as age, health conditions, location, and time work together to contribute to the overall trend in deaths.

Despite data quality issues, insights emerged, including regional disparities, age-based vulnerability, and the impact of contributing conditions. This highlights the importance of understanding these factors for effective pandemic response, and overall emphasizing the need for more reliable and consistent data.

**Introduction:**

The COVID 19 pandemic struck the globe on an unprecedented scale for the 21st century. After its initial outbreak in China, it quickly spread across the globe, along with fear and uncertainty. As time progressed, it found its way into the US, where it eventuated to produce the second highest per-capita death toll of any nation. We investigated CDC data to find trends, patterns, and insights that reveal potential reasons that factored into the end of over 1,000,000 American lives. The CDC dataset we used includes data for each state about its monthly death count, contributing conditions, and age groups, from January 2020 up to September 2023. We also looked at vaccination data from the CDC. The CDC datasets presented some issues that we will discuss further.

**Methodology:**

We conducted our analysis in Tableau and Python, creating visualizations and dashboards that show aspects of the data surrounding death, health conditions, and vaccination. Upon delving into the CDC datasets, we found a plethora of inconsistencies that needed to be addressed. We found issues in the total values for variables such as ‘state’ and ‘age groups’, where the provided numerical totals did not match actual calculated sums. Upon discovery of these inconsistencies, we chose to rely on calculated totals for our analysis. We also found that the ‘State’ column had 54 total unique values, including counts for the District of Colombia, New York City, Puerto Rico, and the United States. The total death count value for the United States did not match calculated sums of each state. Furthermore, the structure of the dataset meant that many of the datapoints were repeated with slight categorical changes. This meant that we had to filter the data to prevent inclusion of duplicates that would overrepresent our total death counts.

For our first dashboard, we looked at the CDC’s COVID deaths by ending week date and state. Our final dashboard is designed to show the how death rates evolved over time across different states and age groups. We created an interactive dashboard that allowed users to filter the data by time, age group, and state, effectively showing how the outbreak evolved through certain populations and regions. This dashboard consists of 3 visuals and a time filter.

The first visual consists of both monthly and cumulative death counts, filtered by all other aspects of the dashboard, such as age group and state. We created a continuous date field called ‘combined date’ using the ‘Month’ and ‘Year’ columns of this dataset, allowing us to see how deaths tracked over time. This field also acts as an interactive filter, allowing the user to select specific time intervals between January 2020 and September 2023. The total death count area graph on this dual-axis visual is achieved using a table calculation, computing the running total of COVID deaths over the time span. Tooltips are also edited to show the monthly and cumulative death toll over the selected interval, regardless of which chart the user’s mouse is hovering, ensuring ease of access to both key values regardless.

The second visual is a map of per-capita covid deaths for each state across the U.S. Darker states represent higher per-capita death counts, whereas lighter states represent lesser death counts. This map acts as a filter, updating the line graph as well as the age distribution graphs to show the values for each selected state. A calculated field was used to create the per-capita death toll, using population data integrated from the 2020 census. This field was applied to the color feature of the map, revealing how death densities were distributed throughout the country. A level of detail expression was used to calculate the difference between state per-capita death tolls and the national average, further showing how certain populations reacted to the virus. This calculation was added to the tooltip over each state, allowing the user to see how state death rates totaled relative to the national average.

The third graph shows the total COVID-19 deaths per age group, highlighting how older people were significantly more at-risk for dying at the hands of the virus. This graph also acted as a filter, allowing the user to section off certain age groups to see how they were spread across each state and across time. From this graph, we can see that the older population exhibited dramatically higher death rates.

For the next dashboard, we analyzed contributing conditions across age groups. The CDC’s Contributive Condition’s dataset is also inconsistent, having the same issues with the others. Several filters were put in place to eliminate duplicates across variables such as ‘State,’ ‘Age group,’ and ‘Group,’ where total values were included alongside individual groupings.

Null values in month and year had to be filtered out, along with flags indicating missing data or ages that were unspecified.

After filtering out any values that would skew the data, we begun working on visualizations. To show the impact Condition groups on COVID-19 deaths, we chose a packed bubble chart that allowed the user to filter the data by condition. To represent the different age groups and how many COVID-19 deaths they have, we chose a simple bar graph, which also acts as a filter for the condition and death counts. We also included a horizontal bar graph to show how distinct conditions contributed to total deaths.

Our third dashboard was originally planned to look into COVID-19 death rates among different racial groups. At first glance, the data, backed by some online research, seemed to tell a clear story: minority groups, particularly Black and Hispanic individuals, faced significantly higher death rates. There was even a point where Black individuals were three times more likely to die from COVID-19 than their White counterparts. But as we kept track of the data over time, we noticed a shift. The death rate among Black individuals started to drop, while the rate for White individuals remained the same, eventually surpassing the former. This intriguing trend shift as well as our initial findings made us want to dig deeper into the role race played in COVID-19 deaths and see if there were other patterns that other people had not analyzed online.

Our goal was to pack this dashboard with various visualizations to explore these racial trends more thoroughly. We wanted to look at distributions, account for population differences, and maybe even look for a correlation between race and COVID-19 death rates. The data was all there for us, but turning it into clear visuals turned out to be a lot harder than expected.

The struggle began with trying to modify our dataset, named 'Provisional\_COVID-19\_Deaths\_\_Distribution\_of\_Deaths\_by\_Race\_and\_Hispanic\_Origin', into something Tableau could work with. Tableau likes its data in certain formats, especially when it comes to summing or averaging. Our dataset, however, was not Tableau friendly. It had separate columns for each race and this 'Indicator' column gave us either death counts or population distributions. We tried to use the data as is, even cleaning it up in a Jupyter Notebook by removing extra unnecessary columns and focusing on the monthly data. But no matter how much we tweaked and twisted it in Tableau, it just wouldn't work. The data needed to be modified, with races needing to be under a single 'Race' column, and those 'Indicator' values needing their own separate columns. We did manage to pivot the race columns in Tableau using the Pivot feathre, but we couldn’t do the same for the 'Indicator' column.

We even tried to manually rearrange the dataset in Jupyter, seeking help from ChatGPT and searching the internet, but still could not create a tableau working dataset. So, we had to let go of this dashboard idea and move on to our next dataset, 'Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age'. Unfortunately, once again this dataset had its own problems. The 'Age Group' column was a mess of overlapping intervals, making any solid analysis not possible since data was overlapping everywhere.

So, we fell back on what we could work with, which was the data we already had. Our third dashboard morphed into a mix of our first two, analyzing 'Conditions\_Contributing\_to\_COVID-19\_Deaths\_\_by\_State\_and\_Age\_\_Provisional\_2020-2023'. We cleaned it up with Jupyter Notebook to align with our other data. This interactive dashboard combines elements from our first two dashboards, allowing users to filter data by time, age group, and state, as well as by Condition Group. Additionally, we've introduced a new feature that enables filtering by U.S. regions: Midwest, Northeast, South, and West.

The standout feature of this dashboard is the line graph. Its primary function is to illustrate the effects of various filters on COVID-19 death rates, enabling the analysis of potential trends such as seasonal death surges and data outliers. To enhance the graph's utility, we developed a 'Seasons Field' using Python, which classifies each data entry by season (Winter, Summer, Fall, Spring) based on the month. Furthermore, to identify outliers, we implemented a parameter for selecting the number of standard deviations, calculated fields for determining the data's lower and upper bounds, and another calculated field to highlight data points that fall outside these bounds.

The other noticeable difference in this dashboard is the regional filter within the map visualization. By integrating Python into Tableau, we managed to classify each data row by its corresponding U.S. region. This enhancement allows for a better analysis of trends and patterns across different regions of the United States.

Despite exploring various other visualization options, we encountered limitations. For example, we considered examining correlations between conditions and death counts but faced challenges due to the mix of categorical and numerical variables. Our attempts to incorporate machine learning for predicting COVID-19 deaths was also hindered by a lack of features and data inconsistencies such as significant decrease in death counts over time. Perhaps if we implemented a Long Short-Term Memory (LSTM) model, we could come up with something significant; however, this was out of our expertise. Also, we tried to make a forecasting visualization; however, all the forecasted values were the exact same and there were no meaningful forecasted values throughout the future months.

In summary, this dashboard is a unity of our previous ones, tailored to uncover more intricate COVID-19 death trends and outliers using an expanded set of variables. Users can now conduct a more detailed analysis, such as comparing COVID-19 death rates among different demographics or regions, conditions, or age groups. The goal is to provide greater flexibility and depth in analyzing the available data.

The last dashboard covered deals with the correlation of vaccine and death rates. This dashboard consists of two graphs and multiple features. The first graph consisted of a line chart that had two variables running through it; people vaccinated, represented with a black line, and covid deaths, represented by the blue filled line. It is important to note that the people vaccinated represents all people vaccinated at least once but does not duplicate a person in the data if they get more than one dose. In simpler words, once a person gets vaccinated once, they will not be added to the vaccination graph again. It is good to note that in order to have the real number of vaccinations, we had to add a calculated field due to get the real number of vaccines given an instance. With these two variables, we overlined the blue area data with the black line to show how the vaccine rate correlated with the covid death rate from December 2020 until May 2023. These dates were prevalent in our data research and give a great outline as to when COVID was important. This shows the data not only for the United States, but for each state when a state is clicked on in the map graph.

Speaking of the map graph, this was the second/final graph of the visualization. With this, we’re showing the vaccinations per one hundred people by state; dark blue representing a high vaccinations per one hundred people and a light yellowish blue representing the lower side. While the map graph gives a simple way of seeing correlation for certain regions in the map, we also gave it the feature of filtering data through the line chart. In other words, if you click a certain state, that state’s data will show up in the line chart.

As well as the line chart showing data for a certain state, there is a number to the right of the map showing the total amount of people vaccinated in the United States. This number will change based on which state is selected in the map.

**Results**

The biggest thing we learned is that the CDC is bad at maintaining good data. Overall, numerical values were inconsistent between files, as well as within them. Each of the datasets we looked at had totals that did not amount (for instance, the values for ‘Male’ and ‘Female’ did not add up to equal ‘All sexes’ in the ‘Sex’ column). This was found to be true for each variable that was split into different groups. Each one of these variables had a total value that was inconsistent with the true sum of the datapoints. Furthermore, the entirety of the data was split between multiple .csv files with different variables, timeframes, and group divisions (for instance, in the “*Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age.csv*” file, the age groups were split into groups of between 5- and 15-year differences, with all age groups exhibiting some kind of overlap, while the age groups in “*Provisional\_COVID-19\_Death\_Counts\_by\_Week\_Ending\_Date\_and\_State.csv”* were consistently split in groups of 10 years*.* These inconsistencies made corroborating our dataset considerably more difficult, as not only did totals not match up, but the variables themselves didn’t either. These inconsistencies led us to analyze the datasets more individually in order to keep discrepancies more contained. However, despite the poor data quality, we were still able to gather interesting and profound insights.

For instance, upon analyzing state-wide per-capita covid trends, we found that southeastern states had a higher per-capita death toll, with the highest states being Mississippi, Oklahoma, West Virginia, and Kentucky. These states also exhibited low per-capita vaccinations, indicating that vaccines are successful in mitigating deaths. Furthermore, states where per-captia death rates were low, such as the New English states, we saw high vaccinations.

Another insight we found was that COVID disproportionately killed older members of the population at a significantly higher level than younger members. According to this data, 90.74% of COVID victims were over the age of 55.

We also found many conditions that contributed to the death count. While the many of COVID-19 deaths listed COVID-19 as a contributing factor, there were more COVID deaths that listed Respiratory diseases as a primary cause of death, such as pneumonia and influenza, followed by circulatory diseases like heart disease and cardiac arrest.

While looking at how conditions contribute to deaths across each age group, it becomes clear that some conditions have a larger effect on particular ages. People 85+ are more likely to die of circulatory diseases, while 75–84-year-olds die of respiratory conditions, while 55–64-year-olds die of obesity. Conditions like dementia and Alzheimer’s only affect the older populations, for there are no deaths for anyone under the age of 65.

Looking into the Covid Deaths vs Vaccines Dashboard, we found that over time, vaccines and covid deaths correlated in an interesting way. We first see that the graph starts with a peak in covid deaths and then a few months later has a peak in vaccines. This wasn’t the only time we saw this kind of connection in the data, as the next time covid deaths peaked, vaccines once again were on the rise months later. Covid deaths rose again shortly after the second highest vaccine peak and there wasn’t a rise in a person's vaccination status this time. There would, however, be a small uptick in vaccinations in the winter of 2022.

We also looked at all fifty states on a map and colored accordingly what each state’s “vaccinations per 100 people” looked like. We saw that the northeast (New England) was mainly dark blue (indicating a high vaccination per capita) and that the southeast was a lighter yellow (indicating a low vaccination per capita). We also saw that for the most part, each state over time kept a similar distribution in how the whole United States was distributed with vaccinations and COVID deaths.

An interesting note to the previous finding in the vaccinations map was how some states (such as Connecticut and Massachusetts) had a value that was over one hundred, meaning, they had more people vaccinated than they had people in their state. The data was funky in multiple aspects like this but overall, we get the idea that the northeast correlates with low death rates.

**Discussion**

It is clear that, as the virus begun to spread, it infected and killed more old people, likely because their older age presented more health risks. Another key insight is that many southeastern states exhibited higher per-capita death rates. This may be indicative of lower socioeconomic conditions, as well as vaccine hesitance, government distrust, and no official standardized and enforced lockdown protocol. Other factors that could impact these numbers may be the distribution of conditions across the U.S., for perhaps higher concentrations of contributing conditions may exist within certain geographic boundaries. Contributing conditions are key in understanding the total death toll across the U.S.

One look reveals the main contributing conditions to COVID-19 deaths, but it takes deeper thought to figure out why. Respiratory diseases are understandable as to why they contributed the most to COVID-19 deaths, as patients who develop comorbidities have a harder time fighting them off. On the other side of the spectrum, Circulatory conditions may have contributed so much as well as COVID-19 since fighting two diseases on two different fronts is very difficult. Interestingly, the number of COVID deaths that are due to respiratory diseases is technically higher than the total COVID deaths due to COVID. This may be due to the way that cause of death is reported, as many COVID deaths are actually manifestations of other conditions such as pneumonia that coincide alongside COVID-19.

Different contributing conditions affect different age groups, so that tracks with the death counts of different ages due to different condition groups. The older the patient, the more susceptible they are conditions, such as heart conditions or lung disease, while other younger patients are more susceptible to developmental conditions such as obesity.

Going back to our last dashboard, we believe the connection previously explained with the peaks in vaccinations happening months after COVID deaths happened is based on people having an increased awareness of the situation. Whether it’s a family member, a friend, or just the news having an impact, we figure that each time a covid death happens, a person is more likely to get vaccinated and protect themself from the virus. As for the third rise in vaccinations that didn't come from a huge rise in COVID deaths, we mainly can attribute this to how most people were already vaccinated up to this point. So we can assume that this rise in people vaccinated probably consists of the last batch of people who were serious about getting vaccinated for protection.

Now about the map graph, with this, we concluded that the northeast was getting vaccinated in high rates while the southeast was not. This correlates with the death per capita map in that, the northeast had the lowest death rate while the southeast had some of the highest rates of death. In simpler terms, the darker the graph, the more people that are vaccinated in these areas which makes correlating areas and data an easier process.

**Conclusion**

The examination of nationwide COVID-19 data through rigorous analysis and visualization techniques has revealed critical insights despite challenges posed by data inconsistencies within CDC datasets. Our investigation highlighted regional disparities, age-based vulnerabilities, and the significant impact of contributing conditions on COVID-19 outcomes.

The observed higher per-capita death rates in southeastern states suggest a confluence of socioeconomic factors, vaccine hesitancy, and governance issues contributing to increased vulnerability. Additionally, the stark correlation between age and vulnerability underscores the need for tailored interventions for different age groups.

Contributing conditions, particularly respiratory and circulatory diseases, emerged as key factors influencing COVID-19 mortality across varying age brackets, emphasizing the complexity of health interactions in pandemic outcomes.

Despite data quality issues, these insights underscore the urgency of addressing data inconsistencies and enhancing data quality within CDC repositories. Accurate, reliable data is imperative for informed decision-making and effective policy implementation in managing and mitigating the impacts of future health crises.

This analysis not only sheds light on critical facets of the COVID-19 pandemic but also serves as a clarion call for improved data hygiene and standardized reporting practices to bolster our preparedness and response to public health emergencies.

**References**

“Covid-19 Data from the National Center for Health Statistics.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 12 Sept. 2022, [www.cdc.gov/nchs/covid19/index.htm](http://www.cdc.gov/nchs/covid19/index.htm).

“CDC Database.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, data.cdc.gov/. Accessed 12 Dec. 2023.

Rodes, Lucas. “Data on Covid-19 (Coronavirus) Vaccinations by Our World in Data.” *GitHub*, github.com/owid/covid-19-data/blob/master/public/data/vaccinations/README.md. Accessed 12 Dec. 2023.