Original Article

Comparing COVID-19 Among Some Highly Affected States in the United States of America

Christopher M. Sabillon, Christina I. Guzman, B. M. Golam Kibria*

Department of Mathematics and Statistics, Florida International University, Miami, USA

ARTICLE INFO

ABSTRACT

Introduction: Coronavirus disease 2019 (COVID-19), a respiratory disease caused by the coronavirus SARS-CoV-2, has had an immense impact on a variety of sectors both worldwide and nationwide. Vast differences are observed among states within the United States of America in terms of COVID-19 cases and deaths.

Objective: The objective of this paper is to present a means through which we can compare deaths between multiple states, using the index date approach applied by Middelburg and Rosendaal.

Materials and Methods: Using the CDC COVID-19 tracker, we created two sets of ten states focusing on states with (1) the highest number of deaths and (2) the highest number of deaths per 100,000. We applied features of the authors’ technique in order to compare deaths between certain states through visualizations. We referred to the cumulative number of deaths on each day from January 21st, 2020 to September 30th, 2020, as a percentage of the cumulative deaths 40 days after the first death.

Results and Discussion: Comparability was established by synchronizing each state to a baseline date, which allows us to adjust for issues that arise from the scales used within a standard cumulative deaths graph, such as a tendency to be driven by the states with the highest cumulative number of deaths. This technique also normalized each state to a standard start date.

Conclusion: This paper shows the application of a technique that allows for clearer comparisons of COVID-19 related deaths between states, as opposed to the use of a standard cumulative deaths graphs.

Introduction

Coronavirus disease 2019 (COVID-19), a respiratory disease caused by the coronavirus SARS-CoV-2, is linked with a vast assortment of symptoms, that can start off as mild as coughing or congestion and develop into more severe symptoms such as shortness of breath, fever, and gastrointestinal distress (4). Starting out as an illness described as “pneumonia of an unknown cause,” COVID-19 quickly spread throughout the world after it was first reported in Wuhan, China on December 31st, 2019 (24). Approximately 9 days after this first report, officials in China determined that this illness was caused by a novel coronavirus, eventually termed COVID-19 on February 11th, 2020 (24). Within a month of the first reported case, instances of COVID-19 were documented in numerous regions worldwide. COVID-19 was first reported in the United States (USA) on January 21st, 2020, with reported cases seen 3 days later in Europe and 8 days later in the Middle East.

* Corresponding Author: kibriag@fiu.edu
By March 11th, the extent of the spread of COVID-19 led it to officially be declared a pandemic by the World Health Organization (WHO) (24). As a response to the pandemic, states within the USA implemented different measures in order to control the virus, such as intermittent lockdowns, enforcing social distancing measures, and requesting the use of personal protective equipment. As a result of COVID-19, it is estimated that there are approximately 33 million cases and over 1 million deaths globally, with the USA leading the world in both cases (approx. 7.3 million) and deaths (approx. 208,000) as of September 30th, 2020 (10). In the USA, Black or African American, non-Hispanic individuals (2.1 times) and American Indian or Alaska Native, non-Hispanic individuals (1.4 times) are more likely than White, non-Hispanic, individuals to die from COVID-19 (3). When compared with those within the 18 - 29 age range, those who are 75 - 84 years old (220 times) or over 85 years old (630 times) are more likely to die as a result of contracting COVID-19 (2). Certain factions of the population, such as the groups previously mentioned, are not only at higher risk of experiencing the more severe symptoms associated with COVID-19 that may lead to death, but also more susceptible to long-term health issues (12). Various studies are being conducted in order to pinpoint what these long-term effects are, as well as to ascertain the degree to which COVID-19 has an effect on the different systems of the human body. For example, researchers such as Chen et al. (5), Guzik et al. (8), and Puntmann et al. (20) are attempting to evaluate the extent to which COVID-19 affects the cardiovascular system. Work conducted by Fotuhi et al. (7) and Pero et al. (18) focus on determining the neurological complications associated with some cases of COVID-19, such as “stroke” and “encephalopathy.” Research on lung abnormalities associated with COVID-19 is currently being conducted through exploratory means such as chest CTs (19) and autopsies (13) on recovered COVID-19 patients and those who have died from COVID-19, respectively. Studies are also being conducted in hopes of exploring the adverse mental health outcomes associated with COVID-19. Yao et al. examines the extent to which COVID-19 has exacerbated pre-existing mental health conditions, such as anxiety and depression, among individuals, making these disorders a part of a “parallel epidemic” (25). The socioeconomic impact that COVID-19 has had within the USA and other countries is also being assessed by researchers such as Nicola et al., whose work sheds light on the effects that COVID-19 has had on a variety of sectors, including agriculture, manufacturing, education, etc. (15). A vast amount of research is currently focusing on the development of mathematical models and methods in order to predict the number of COVID-19 related cases and deaths, and ultimately make comparisons between different regions. For example, Singhal et al. use both a parametric approach and a non-parametric approach that relies on the “Fourier decomposition method” in order to both capture the trend of and predict the number of COVID-19 cases (21). Khafaie and Rahim attempt to use the case fatality rates and recovery rates of countries, such as China, Italy, South Korea, and the USA, in an attempt to “identify high risk areas” (11). Research being conducted by the Institute for Health Metrics and Evaluations (IHME), has led to the development of multiple models in order to explore the “possible trajectories of SARS-CoV-2 infections” both on a national and international level, allowing them to make comparisons between the regions (9). Middelburg and Rosendaal refer to the use of a reference date after the first death within multiple countries in an attempt to establish a method of comparison between them (14).
This unique application of comparing countries has inspired the focus of our paper, which draws on their technique in order to establish comparisons between COVID-19 related deaths in states within the USA. In this paper, we will implement Middelburg and Rosendaal’s method as a means of making comparisons between select USA states (14). These comparisons will look at two sets of ten states, including the top ten states with the: (1) highest number of overall deaths, (2) and highest death rate per 100,000 (1). These two sets were chosen based on lists created by the Centers for Disease Control (CDC) as a means of highlighting the applicability of this technique when attempting to analyze the states that had the most deaths as a result of this pandemic. This paper is organized as follows: We describe our data sources and provide some descriptive statistics in the Material and Methods section. We give an in-depth outline of the techniques we applied, as well as visual representations of state comparisons, in the Results and Discussion section. This paper will end with some concluding remarks in the Conclusions section.

Materials and Methods

We extracted our data from The New York Times’ (NYT) public GitHub repository and used it as the main data source for this paper. In an attempt to provide up-to-date and comprehensive COVID-19 data, the NYT’s GitHub repository houses data from sources such as “local governments and health departments” that are being refreshed on a daily basis (22). Their data also reflects any changes made as a result of government reporting decisions, delayed reporting, or justified corrections (22). The dataset from this repository that we utilized contains information about the number of COVID-19 related deaths and cases at the state-level. The NYT dataset includes cases and deaths based on two categories: confirmed and probable. The inclusion of both confirmed and probable definitions of cases and deaths adheres to the advice of the Council of State and Territorial Epidemiologists (CSTE), which is now supported by the CDC (6).

We will be focusing on the number of deaths that take place between January 21st, 2020 to September 30th, 2020. The earliest date in our dataset is January 21st, 2020, reflecting the first reported COVID-19 case within the United State. Aside from this repository, we also referred to the CDC website in order to pick two sets of ten states to compare based on two distinct death-related metrics. The states chosen reflect the CDC’s data as of September 30th, 2020 and specifically focuses on 16 states within the USA. It is important to note that we have included the District of Columbia within our set 2 considering that it is a city that is independent of the states, as it is the capital of the USA. The states and their corresponding set are as follows: (1) highest number of overall deaths: Florida, New York, New Jersey, California, Massachusetts, Illinois, Pennsylvania, Michigan, Texas, and Georgia; and (2) highest death rate per 100,000: New York, New Jersey, District of Columbia, Massachusetts, Rhode Island, Connecticut, Louisiana, Mississippi, Arizona, and Michigan. Note that the states that are outlined for each set are in no particular order. Table (1) provides a state-level breakdown for the 16 states by showing each state’s estimated 2019 population in accordance with the USA Census Bureau, as well as cumulative number of cases, case rate per 100,000, cumulative number of deaths, and death rate per 100,000, all as of September 30th, 2020 (23).
As specified in Table (1), California has had the highest number of cumulative COVID-19 cases as of September 30th, 2020, yet it is not one of the states with the highest cases per 100,000. This indicates that though COVID-19 has affected a substantial number of individuals, these individuals constitute a fraction of California’s total population. Florida, the third state with the highest cases per 100,000 in our list, is not a member of the set with the highest deaths per 100,000 (set 2). It is interesting to note that while Florida is one of the states with the highest cases per 100,000 in the USA, a grand majority of the cases did not translate into deaths when accounting for the size of the population (see Figure (1)). New York, one of the most affected and populous states within the USA, surpasses the state with the second highest number of cumulative deaths (New Jersey) by double the amount. However, the data indicates that New Jersey has a higher death rate per 100,000 than New York.

Figure (1) provides a daily breakdown of cases and deaths in Florida from March 1st, 2020 (date of first two reported cases in Florida), as well as in Mississippi from March 11th, 2020 to September 30th, 2020. We chose to create these graphs as a means of illustrating the number of cases and deaths within one of the most and least populous states on our list of 16 states. While both Florida and Mississippi appear to be following a similar trend in terms of daily cases, these trends lie on very distinctly different scales, with the peak number of cases in Florida being 15,300 on July 12th, 2020 and in Mississippi being 1,775 on July 30th, 2020. The same can be observed when looking at daily deaths, with the peak number of deaths in Florida being 276 on August 11th, 2020 and in Mississippi being 67 on August 25th, 2020.

When looking at these graphs in Figure (1), it is difficult to make direct comparisons between both states, as they had their first cases reported on different days. Also, if we were to have created the graphs using the

<table>
<thead>
<tr>
<th>State</th>
<th>2019 Population</th>
<th>Cumulative Cases</th>
<th>Cases Per 100,000</th>
<th>Cumulative Deaths</th>
<th>Deaths Per 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arizona</td>
<td>7,278,717</td>
<td>218,511</td>
<td>3,002</td>
<td>5,654</td>
<td>78</td>
</tr>
<tr>
<td>California</td>
<td>39,512,223</td>
<td>821,125</td>
<td>2,078</td>
<td>15,898</td>
<td>40</td>
</tr>
<tr>
<td>Connecticut</td>
<td>3,565,287</td>
<td>57,550</td>
<td>1,614</td>
<td>4,508</td>
<td>126</td>
</tr>
<tr>
<td>District of Columbia</td>
<td>705,749</td>
<td>15,326</td>
<td>2,172</td>
<td>627</td>
<td>89</td>
</tr>
<tr>
<td>Florida</td>
<td>21,477,737</td>
<td>706,508</td>
<td>3,289</td>
<td>14,316</td>
<td>67</td>
</tr>
<tr>
<td>Georgia</td>
<td>10,617,423</td>
<td>301,485</td>
<td>2,840</td>
<td>6,861</td>
<td>65</td>
</tr>
<tr>
<td>Illinois</td>
<td>12,671,821</td>
<td>296,935</td>
<td>2,343</td>
<td>8,935</td>
<td>71</td>
</tr>
<tr>
<td>Louisiana</td>
<td>4,648,794</td>
<td>167,458</td>
<td>3,602</td>
<td>5,511</td>
<td>119</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>6,892,503</td>
<td>132,116</td>
<td>1,917</td>
<td>9,456</td>
<td>137</td>
</tr>
<tr>
<td>Michigan</td>
<td>9,986,857</td>
<td>138,074</td>
<td>1,383</td>
<td>7,084</td>
<td>71</td>
</tr>
<tr>
<td>Mississippi</td>
<td>2,976,149</td>
<td>98,190</td>
<td>3,299</td>
<td>2,969</td>
<td>100</td>
</tr>
<tr>
<td>New Jersey</td>
<td>8,882,190</td>
<td>207,062</td>
<td>2,331</td>
<td>16,122</td>
<td>182</td>
</tr>
<tr>
<td>New York</td>
<td>19,453,561</td>
<td>463,369</td>
<td>2,382</td>
<td>32,757</td>
<td>168</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>12,801,989</td>
<td>164,135</td>
<td>1,282</td>
<td>8,216</td>
<td>64</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>1,059,361</td>
<td>24,748</td>
<td>2,336</td>
<td>1,114</td>
<td>105</td>
</tr>
<tr>
<td>Texas</td>
<td>28,995,881</td>
<td>784,027</td>
<td>2,704</td>
<td>16,102</td>
<td>56</td>
</tr>
</tbody>
</table>

* Note: The COVID-19 related information found above was obtained from the NYT GitHub repository, which may not reflect the data found in other sources.
same scale, our ability to detect the fluctuations in cases and deaths for the graphs corresponding to Mississippi would be very difficult. With the method proposed by Middelburg and Rosendaal, both states, or all of the states, would be synchronized by the same date, the date of the first reported death for each state (as day one) (14). This method would also enable us to place the two states (or more) on the same scale, by referring to the cumulative deaths as a percentage of cumulative deaths on day 40, which allows for a more direct comparison. By applying this technique, we aim to fulfill our objective, which is to compare deaths between certain USA states through our visualizations.

**Results and Discussion**

According to Middelburg and Rosendaal, an alternate approach to establish comparability of COVID-19 related deaths or cases between different regions is to “normalize” the number of deaths or cases to a “reference number,” or date (14). Within their paper, they attempted four methods to make comparisons between cases and deaths within different countries in order to decide which approach led to the clearest comparisons between countries. The results associated with each method led them to conclude that the fourth method allowed for the clearest comparisons. The fourth approach used the date of the first death as day one and then compared the cumulative number of deaths as a percentage of cumulative deaths at day 25 after the first death. This allowed them to later make isolated comparisons between China and other countries.

We decided to apply this unique approach to our two sets of ten USA states in an attempt to establish comparability and perform an in-depth analysis of the trends observed between states. Our approach for determining which reference date after the first death led to the clearest comparison of our sets of states was subjective in nature. After exploring the
NYT dataset, we decided that there was an insufficient number of cumulative deaths on any of the dates between the first date and the 25th date after the first death took place for each of the states. Using any of the dates within that range as our reference date would also lead to comparisons that were heavily based on the first few months of the COVID-19 pandemic in the USA. This can be observed when looking at the dates of the first death in the 16 states, which all took place within the month of March. We decided to observe the number of cumulative deaths in intervals of 5 days after the first death, starting with 30 days. The variability between the cumulative number of deaths for each of the states at 30 and 35 days after the first death is too wide in order to make any valid comparisons. When looking at the intervals that follow (i.e. 40 to 60 days after the first death), we noticed that the graphs began to look very similar with only a limited number of fluctuations. From 40 to 60 days after the first death it appears that overall, the graphs being produced were very similar in both sets of states. We chose 40 days after the first death as our reference date in order to be able to compare as much data as possible. For all 16 of our states within the 2 different sets, the 40th day after the first death lies around late April and early May.

In order to further illustrate the comparability afforded through the unique technique developed by Middelburg and Rosendaal, we created two graphs Figure (2) representing the top ten states with the highest cumulative number of deaths (set 1) and two graphs Figure (3) representing the top ten states with the highest death rate per 100,000 (set 2) (14). The left graphs within both figures show deaths as the cumulative number of deaths by date for 2020, while those on the right show deaths as the percentage of cumulative number of deaths with respect to the 40th day after the first death. The graph on the left of Figure (2) (Graph I) shows a monthly breakdown of the cumulative number of deaths for the states in set 1. The graph on the right (Graph II) compares the same set of states, yet relies on the technique that we have been focusing on as a means of making comparisons between the states. The scale used on the left graph is heavily influenced by the state with the highest cumulative number of deaths, making it difficult to assess the changes in the cumulative increases in deaths for the other states. By comparing the states with regards to a reference date, the 40th day after the first reported death for each state, we are able to
get a better sense of when the fluctuations in deaths took place for all of the states proportional to the 40th day. When specifically looking at the state of New York in graphs I and II, we see that while New York lies on the top of graph I, it lies on the bottom of the graph II. Though New York has consistently had the highest cumulative number of deaths, as indicated by graph I, the increase in the number of deaths remained proportionally consistent following the 40th day after the first death. This indicates that most of the deaths took place prior to the 40th day after the first death and that there was a similar increase in deaths each day following the 40th day. This similar increase in deaths each day is illustrated by a steady, fairly-level, line.

By solely focusing on graph I, we are unable to identify when this consistent increase in deaths began with regards to when deaths began to take place in New York. Graph I allows us to see the cumulative number of deaths that take place on a specific date in 2020, while graph II allows us to reference these cumulative deaths to a synchronized date where they first started, which allows for a more direct comparison. With a synchronized date, each state is compared relative to when COVID-19 deaths began, unlike in graph II, which has different starting points for each state. By establishing the first “day after first reported death” as being our baseline date, we are able to represent all of the states simultaneously. Using the first date that a death was reported in only one of the states as the baseline date makes it difficult to assess how all of the states are faring with regards to deaths after the state began to experience deaths. For example, graph I has a start date of March 10th, 2020 due to COVID-19 death reports beginning on that day in New Jersey, which is earlier than other states on the same graph.

A notable observation can be made in graph II, illustrating an overlapping relationship between New Jersey, Pennsylvania, Georgia, and Illinois until about 125 days after the first death, which eventually diverges. At around that time point it appears that Georgia begins to diverge from the trend that it had been following alongside the other states, whose percentages of cumulative deaths respective to day 40 were very similar. Georgia begins to increase proportionally with respect to day 40, while New York remains consistent, indicating a very slight daily increase in deaths. The similarity seen among the states until day 125 is not necessarily an indicator that these states were experiencing the same number of cumulative deaths up until that time point, but instead an indicator that their cumulative deaths were proportionally similar with respect to the number of deaths that took place on the 40th day after the first death for each state. This proportional similarity is best illustrated by looking at two states at day 125. Georgia (GA) and New Jersey (NJ), which had 767 and 4,070 cumulative deaths, respectively, at day 40, had 2,996 (GA) and 15,525 (NJ) cumulative deaths at day 125, leading them to both be about 385% respective to the 40th day after first death. It is important to acknowledge that there is a difference in the total number of cumulative deaths between these two states on the dates mentioned. Georgia had a significantly lower number of cumulative deaths on their 40th day and their 125th day compared to New Jersey. This can be due to many factors, such as different population sizes, number of resources, or state-level ordinances, which we are not accounting for through the technique used. Notably, on July 1st, 2020, two weeks before the 125th day after the first death (July 14th, 2020) took place in Georgia, Governor Brian P. Kemp’s executive order allowed for the hosting of conventions and re-opening live performance venues (17).
The two graphs (graph I and II) in Figure (3) illustrate the application of the same approach as a means of comparing the top ten states with the highest number of cumulative deaths with respect to the size of their populations. Such as we had observed in graph I of Figure (2), the scale used in graph I of Figure (3) is heavily influenced by the state with the highest death rate per 100,000. This makes it very difficult to look at the cumulative death rates per 100,000 for the District of Columbia and the state of Rhode Island. Though Rhode Island and the District of Columbia are also slightly difficult to see because of their steady, overlapping nature in graph II, we are still able to establish comparisons between the two states and the other states within the graph. As observed on graph II, the two states were initially increasing in a similar fashion to all of the other states except for New Jersey, Mississippi, and Arizona. Around the 60th day after the first death took place for both states, we began to see an increase in the cumulative number of deaths with respect to day 40 followed by a daily increase in deaths all the way up to September 30th, 2020. It is interesting to note that at about day 57 after the first death in Arizona, the stay-at-home order that was put into place as a means of controlling the spread of COVID-19 was lifted. The order was announced by Governor Douglas A. Ducey on May 15th, 2020, or the 57th day after the first death (16). When focusing on the state of New York we are once again able to observe its different position within graph I and II. New York lies on the top of graph I, while lying on the bottom of graph II, indicating that most of the deaths took place prior to the 40th day after the first death and that there was a similar increase in deaths each day following the 40th day. Graph II allows us to compare New York to the other states in terms of how much of the deaths have already taken place with regards to how many have continued to occur up to September 30th, 2020. When focusing on the top two states in graph II, Arizona and Mississippi, it appears that they have had the highest increase in daily deaths with respect to the cumulative deaths on day 40. All of the states aside from these two appear to have fairly consistent increases in deaths daily. We are unable to detect the sharp change in daily deaths that takes place within these two states at around day 60 by solely looking at graph I. All of the visualizations found within this paper, as well as the data cleaning needed prior to their creation, were constructed using
the following R packages: tidyverse, lubridate, ggsci, scales, and kableExtra.

Conclusions
This paper considers the applications of a unique technique (14) as a means of making clear comparisons between deaths in multiple states within the USA. The USA states of interest were chosen based on their high cumulative number and rates of COVID-19 related deaths. We hope that this paper serves as a reference for those who aspire to establish comparisons between regions in order to better understand the impact of COVID-19 and ultimately apply this understanding in a meaningful way. Considering that the effects of COVID-19 can vary greatly on a regional-level (or state-level), the use of a reference date as a means of establishing comparability among states allows us to identify interesting trends that were initially not as clear through standard cumulative death graphs. It is difficult to make any valid or clear comparisons between states when looking at standard cumulative death graphs, due to possible scaling issues or different start dates of COVID-19 reported deaths for each state. The technique applied in this paper synchronizes each state by using the first date of a reported death as day one in order to avoid the technical issues associated with different COVID-19 starting dates. By using the cumulative deaths as a percent of cumulative deaths on day 40, this method places each state on a similar percentage scale, serving as a remedy to the initial scaling issue in standard cumulative death graphs. This is illustrated in Section 3 within Figures (2) and (3) (i.e. New York), when comparing the graphs’ cumulative deaths versus those created using the technique outlined within this paper.

Conflict of Interests
We have no conflicts of interest to disclose.

Acknowledgments
Authors are thankful to the referees for their valuable comments and suggestions, which certainly improved the quality and presentation of the paper. We wish to dedicate this paper to all of those who have been lost or greatly affected by the COVID-19 pandemic.

References


