COVID-19 in SNF (Skilled Nursing Facilities)

John Weisner, Shravan Shenoy, Arie Van Hofwegen.

Author contributions

John Weisner contributed hard coding, data trimming, dataset analysis.

Shravan Shenoy contributed plotting and examining data, dataset analysis.

Arie Van Hofwegen contributed writing project report, dataset analysis.

Abstract

In order to conceptualize the impact of skilled nursing facilities during the COVID-19 pandemic, our team plans to analyze case data from different facilities within the state of California. This analysis will hopefully give rise to new suggestions and goals that we can address as we search for a reachable solution to help both residents and health care workers during the COVID-19 pandemic. In order to do this we will first subset data into relevant facilities with valid resident and worker cases, and then compare the two in hopes to find some kind of trend. In the end, we found that there was significant correlation between the amount of cases present among residents and the amount of cases present among staff. The county and more specifically the facility also played a role in if Covid was transmitted from residents to staff, with some facilities doing a much better job than others.

Introduction

For our project, we aim to examine the impact of the Covid-19 pandemic on skilled nursing facilities in California. This topic deals with a scientific field of study best described as pathogen research and examination. We hope to use this topic in order to better understand the effectiveness of skilled nursing facilities during the COVID-19 pandemic.

Background

COVID-19 is a worldwide pandemic that has affected the lives of billions of people in 2020 and 2021. COVID-19 specifically impacted the skilled nursing facilities, showing an increased risk of both infection and death compared to normal residents. But why is that? This project aims to analyze data taken by the state of California in order to better understand the reason that skilled nursing facilities were either successful or unsuccessful in deterring the spread of COVID-19, and if not, how successful were they at minimizing the death rate. By analyzing the dataset, we could potentially learn why specific skilled nursing facilities are doing a good job at keeping in-house patients safe and how should receive more funding, resources and help, or if not, which facilities should be deprioritized.

<!-- Provide enough background to position the reader to understand your project aims and their relevance. Most likely, you can adapt your background section from the plan report for this section. If you were satisfied with what you wrote previously and no revisions were suggested, you can simply copy that material.

You may want to consider making revisions anyway now that you've carried out your data analysis -- perhaps some pieces of background information aren't as relevant to understanding your work, or some pieces need to be elaborated in further detail. -->

Aims

As we continue to examine the data set we are working with, we have decided to approach the material in a more conservative way. Instead of analyzing all of the data from each skilled nursing facility in California, we have trimmed the data to only nursing facilities that have shown resident cases of greater than 175 at any particular time. This will help us constrict the data to only facilites that have not been particularly successful at minimizing the spread of COVID-19, while also allowing us to examine how successful they were at fixing their mistakes at future dates in the data. Furthermore, the facilities with large amounts of cases provided much more complete data. Since any data values less than 11 were omitted for confidentiality reasons, small facilities had most of their data missing and that is not as useful for analysis. Alongside this, we also want to examine the success of facilities based on the county they belong to. That is, if one county was more successful in preventing the spread of COVID-19 than another, and why. For this project, we had two primary questions. First, how did the amounts of Covid cases change over time and were these similar between different counties. For this, we aggregated Covid data for each county over the period of a month and charted the results, finding that there were several patterns among counties and month. Our second question focused on exploring if there was any correlation between health care worker cases and resident cases. Again for this, we looked at the data split amongst counties and found that there was correlation between the two.

Materials and methods

Datasets

The data is from California Health and Human Services (CHHS). The data is updated daily, so it is very likely collected by the individual nursing facility and supplied to the government for record keeping, especially considering the importance of Covid right now.

The data can be accessed via the following link: https://data.chhs.ca.gov/dataset/test-cdph-skilled-nursing-facilities-covid-19/resource/d4d68f74-9176-4969-9f07-1546d81db5a7

Although it is not explicitly stated what the exact procedures are, it can be assumed that the data is collected and recorded by the staff at the facilities daily and that information is then relayed to this dataset where it is updated.

The population is skilled nursing facilities in California. The site does not go into much detail about the sampling design, but we will treat the sampling frame is a large proportion of the skilled nursing facilities in California, but not quite a census, making it administrative data. The scope of inference of this data extends just to the skilled nursing homes in California that we are focusing on, which is 12 in total. The observational units are the skilled nursing facilities, otherwise known as SNF's.

Provided below is a table of descriptions of the variables used throughout the analysis of the dataset.

Name	Variable description	Type	Units of measurement
County	County of facility location	Text	N/A
facility_id	Nine-digit hospital identifier	Text	N/A
facility_name	Name of the skilled nursing facility for which numbers are being reported	Text	N/A
as_of_date	Date for which counts were reported	Timestamp	Days
total_resident_cases	Number of covid cases among the residents of the facility. After recovery, the previously afflicted individual would no longer be counted	Numeric	Covid Cases
new_resident_cases	Number of new covid cases on the day reported among residents at the facility	Numeric	Covid Cases
total_resident_deaths	Total number of residents who died or are suspected of dying because of Covid since January 1, 2020	Numeric	Covid Deaths
total_health_care_worker_cases	Number of covid cases among the staff of the facility. After recovery, the previously afflicted individual would no longer be counted. Staff will be counted at each location that they work at	Numeric	Covid Cases
new_health_care_worker_cases	Number of new covid cases on the day reported among staff at the facility	Numeric	Covid Cases
total_health_care_worker_deaths	Total number of staff who died or are suspected of dying because of Covid	Numeric	Covid Deaths
note	Tells which columns in the row have missing data, from a to f corresponding to total_resident_cases to total_health_care_workers_deaths	Text	Columns

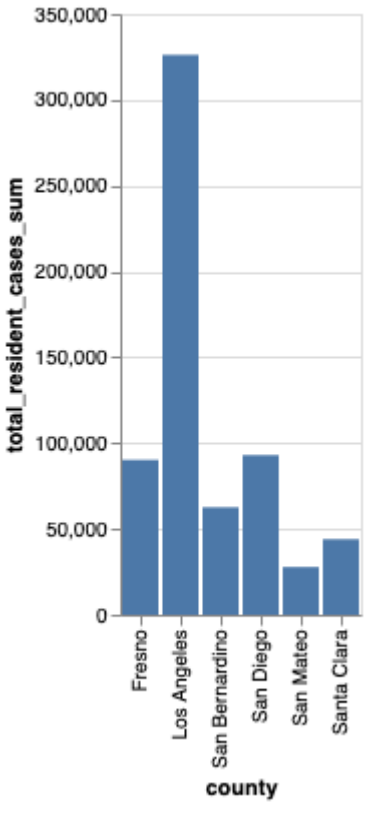
Methods

Compiling and searching for a dataset was done by searching through various databases online. Once a satisfactory dataset regarding county resident covid cases and hospital cases was found, we then skimmed through the dataset to determine what conjectures we can create and test through a thorough analysis. The covid case dataset was then trimmed to include only 12 main facilities across 6 counties, decreasing the number of observations from several thousands to roughly a select hundred. This was done as a considerable amount of the data was missing values due to confidentiality reasons. We adapted the dataset to include data that did not contain missing information for imparative variables. The relationship between specific variables of interest was then plotted. For instance, the total number of resident cases per county across 6 counties was analyzed, allowing us to see which counties were affected the most, and why this could be the case. Afterwards, we theorized and then created plots to understand how total resident cases were affecting health care worker cases, and how this changed over the course of several months through various plots, as presented below. A correlation matrix was then created to answer the conjecture: What is the relationship between the amount of cases present among residents and the amount of cases present among staff, and how strong is the correlation assuming any. From this we were able to draw several conclusions and discuss the results of our analysis.

Results

Figure 1

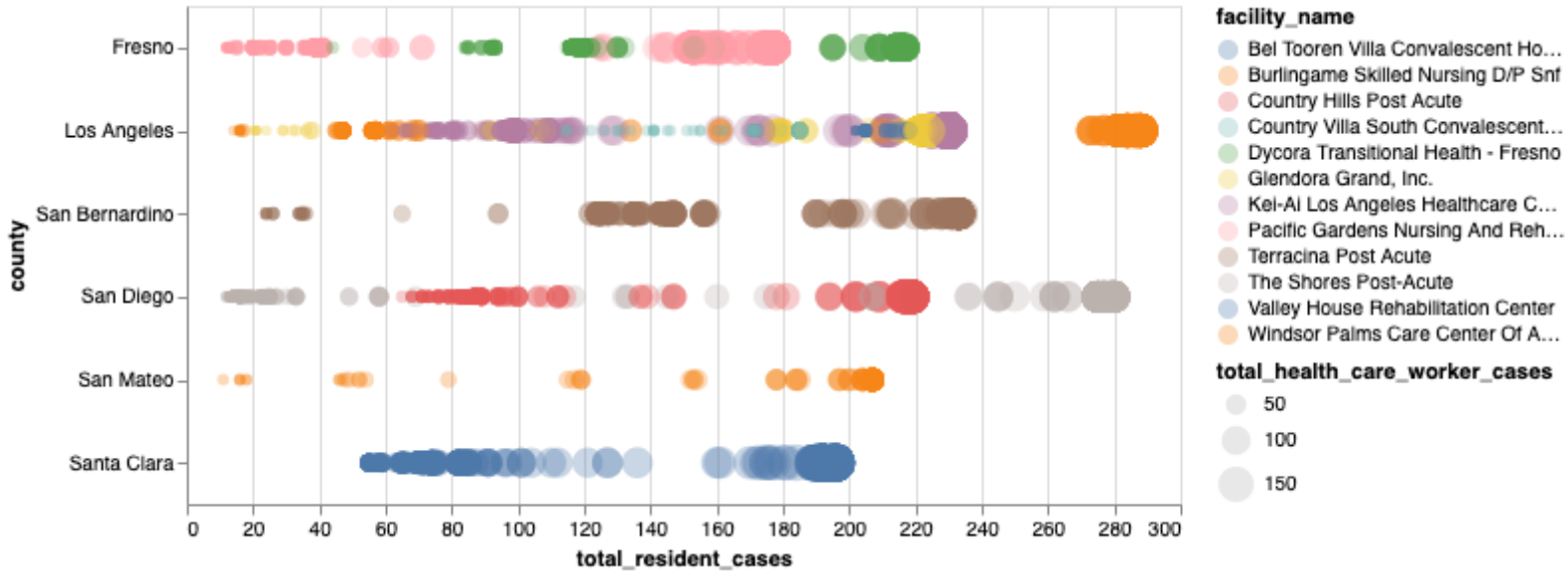
The first plot is a simple bar chart showing how many Covid cases there were among residents in each of the six counties with nursing facilities that we analyzed.



This chart shows that Los Angeles was the most affected of the counties with regards to Covid. This is logical becaue Los Angeles is the most populace of the counties and had the most nursing facilities that fit the criteria for analysis (4 in total). One thing of note is the cases present in Fresno county. Fresno has less population than Santa Clara and San Bernardino and about one-third of the population of San Diego County, but has close to the same amount of cases as San Diego. One thing to note about this graph is that the total cases can double/triple/quadruple/etc count people because everyday that someone is afflicted they are counted again amongst the cases. However, this graph is still useful for analysis because it can show how quickly different counties either cut down on cases or get people healthy. A better stat to use for this graph would be aggregating new cases instead of total cases, but unfortunately many facilities had missing data for new cases because confidentiality only allows new cases to be reported if there are more than 11 in one day, which is not often the case.

Figure 2

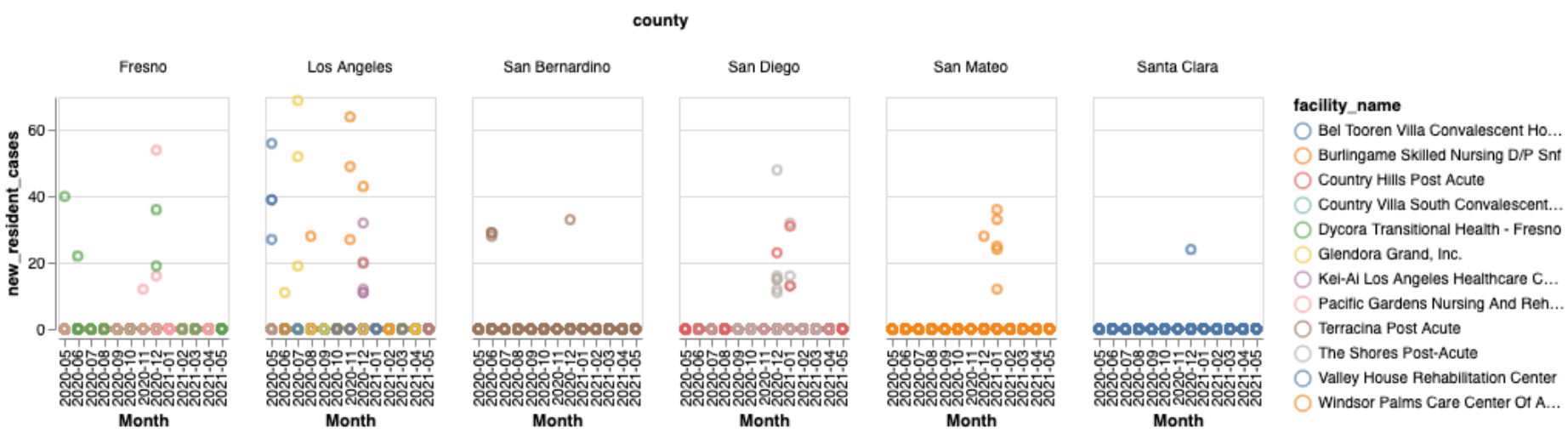
The next figure is a comparison of total resident cases and total health care worker cases among the specific facilities. These comparisons are separated by which county the facility is located.



For all the counties, health care worker Covid cases increase as the total resident cases increase. This is fairly logical because the workers come in contact with the residents a good amount, so more infected afflicted residents should lead to more infected staff members. Among the counties themselves, San Mateo seems to have the best luck in preventing workers from getting infected when there were high amounts of resident cases. In Los Angeles, there seems to be variance in the effectiveness of preventing spread among staff. Some facilities are right along with normal trend of infection, but a few (for example Country Villa) have very low staff rates in comparison to resident rates. This graph has significance because it can show which facilities are doing a good job at preventing staff infections and which are doing a poor job. If there is a difference in how the facilities are carrying out their daily operation, then this chart can help separate the good performing facilities and poor performing facilities so administration knows where to look to update their procedures.

Figure 3

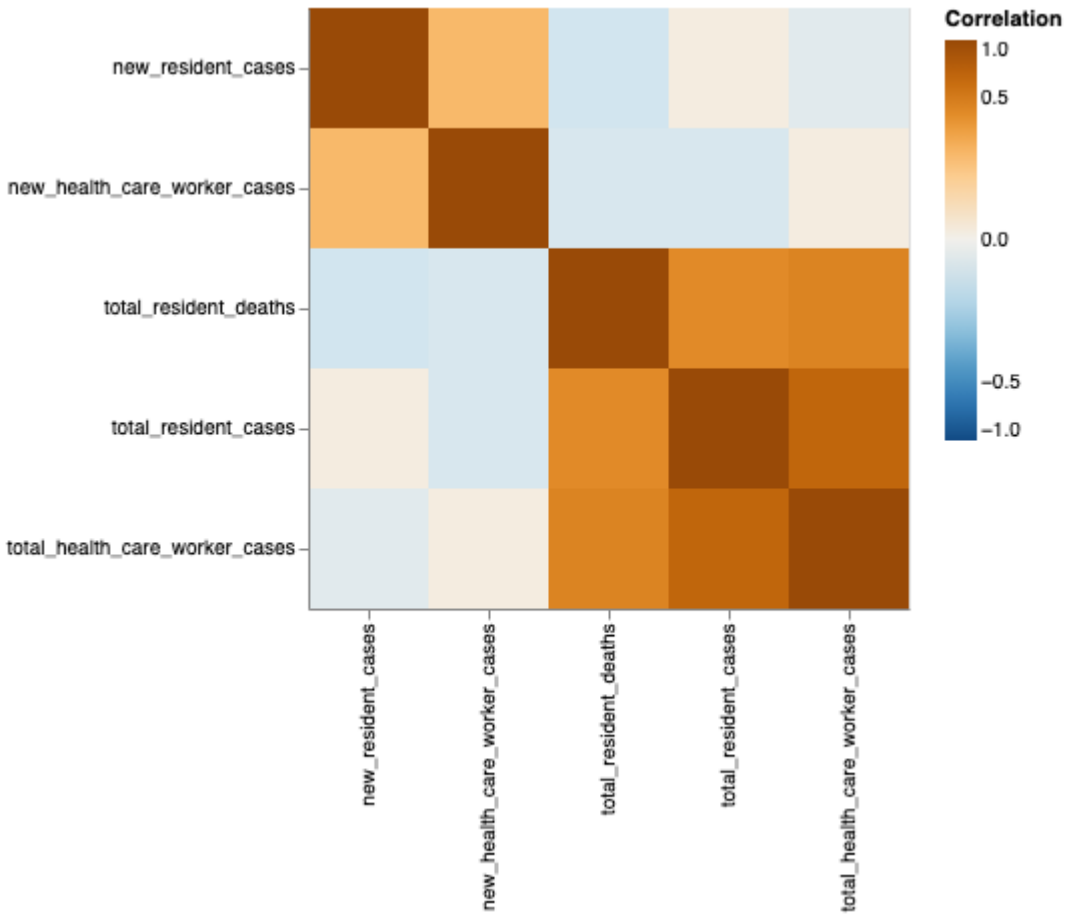
This next chart shows the reporting of new Covid cases for each month among each of the analyzed facilities faceted by county.



This set of graphs shows when new covid cases arrived from each facility. Unfortunately, for many of the facilities there were "0" new cases total in a month due to having to suppress data for confidentiality reasons. However even with the suppressed data, these charts are useful to see where and when the spikes in cases occurred and if the spikes were different among the different counties. Knowing if spikes are different is important because it would help show effectiveness of different Covid protocals in different counties. As shown above, in all counties there was a spike in data in either December of 2020 or January of 2021. Many counties also had large amounts of cases near the beginning of the pandemic in May of 2020.

Figure 4

This next figure is a correlation matrix of 5 different variables: new resident cases, new health worker cases, total resident deaths, total resident cases, and total health worker cases. Total health worker deaths was excluded because there were virtually none present and those that occurred were reported as missing because of confidentiality.



This chart shows the correlation between the various variables, with dark values of brown/blue representing strong positive/negative correlations respectively. Light brown or light blue squares represent correlation values close to 0. The strongest correlations is positive between total_resident_cases and total_health_care_worker_cases with value 0.67. The other two notable correlations are one between total_resident_deaths and total_health_care_worker_cases with value 0.45 and the other between total_resident_cases and total_resident_deaths, with value 0.41. The other correlation values are close to zero. The strong correlation between total_resident_cases and total_health_care_worker_cases makes sense and confirms what is seen in Figure 2, with health worker cases rising as total cases rise. One thing notable about this chart is the low correlations that new_resident_cases and new_health_care_worker cases have with the other variables. This is likely do to the fact that many of that data is missing and I would speculate that these correlation values would be much higher with the other variables if all the data were available and none of it was confidential.

Discussion

From the data we analyzed, we saw that different counties had vastly different experiences with Covid. Los Angeles county was hit the hardest, but that was expected due to its large population. Fresno, on the other hand, was hit hard despite having a relatively small population whereas the nursing facility in San Mateo seemed to handle the virus well. Spikes in Covid were fairly consistent among the counties, with all of the counties receiving an uptick in cases in December 2020 or January 2021. This makes sense in the context of California because Governor Newsome lifted the stay-at-home order in January, likely corresponding to the increase in cases. Additionally, there were many new cases in May of 2020, but then new cases tapered off as time went by, indicating that there was a somewhat successful response and adaptation to Covid after the initial shock. One aim we had in this project was to analyze the relationship between health worker cases and resident cases. We found that there was fairly significant correlation between the two wih a correlation value of 0.67. In most situations, a high amount of patient Covid cases leads to a high amount of health care employee Covid cases.

However, this dataset does have some caveats that come along with it. The scope of inference of these results is limited to the 12 facilities that we analyzed. Smaller facilities may have different ways of stopping the spread of the virus, so our data should not be used to generealize other data. Furthermore, the dataset that we used contained a large amount of missing/confidential data. It would have been nice to further explore the variable relating to new Covid cases, but much of that data is missing. Furthermore, it would be interesting to explore data related to the total amount of workers and patients at these facilities in order to gauge how large and exactly how well they prevented the spread of the virus.