STAT 479 Project Report (Covid)

Shihao Yang Yifan Zhang Kexiao Zhu Anze Xie Sarah Tian

Dongyang Leng

Abstract

In our project, we conducted Bayesian analysis with the data set provided by the Centers for Disease Control and Prevention on November 15, 2021. We used a hierarchical model to figure out how the effect of age groups on the covid-19 infection rate varies across different state in the United States. In this report, we will introduced how we processed the data set and constructed the model. Also, we will list the our results and analyze the reasons for some states having a higher infection rate for a specific age group.

1. Problem Statement

The goal of this project was to detect the relationship between Covid-19 infection rate and different age groups in each state. The project was centered around the question "which state is more likely to have higher infection rate for different age groups?". Specifically, we were interested to reveal infection rate across different age groups geographically and compared the infection rate of given age groups between different states. We also attempted to pinpoint the possible causes of high infection rates in certain states and suggest a possible direction for related research.

2. Motivation and Importance

Covid-19 has largely impacted the economy and people's lives. Since the beginning of the pandemic, public institutions, businesses, and schools have been closed due to the contagiousness of this virus. Also, given that highly effective vaccinations were created and maintained a stable number of infections for a period, new variants of Covid-19 still emerged continuously and new rounds of infection may be expected. To alleviate the effects of Covid-19 on people's lives, it is vital to conduct related research and introduce effective policies. Therefore, the analysis of this project can be useful to convey some information about what age groups may be more susceptible in some particular states. Relationships between infection rates and policies are discussed to provide useful insights for future research and policy making.

3. Data Overview

The original dataset consists of 37.5 million rows and 19 columns, where each row is a case and the information of each case is shown in columns. To better clean the data and filter out the useful information, we converted characters such as "Missing" and "Unknown" to NA and visualized the percentage of missing data within the dataset. Figure 3 shows the percentage of missing values in each variable. The columns that we were interested in, such as res_state and age_group, have relatively low percentage of missing rate, indicating that it was safe to remove the cases with missing data in these two variables. After removing the missing data, only res_state and age_group were selected and grouped into the final dataset, which contained rows of states with different age groups in columns. The population and the sum of infection numbers in each state were also appended to the dataset. Thus, the final dataset consists of fifty rows and seven columns.

4. Models

In this section, we present our model settings and how we fit the model with the COVID-19 Case Surveillance Data we collected.

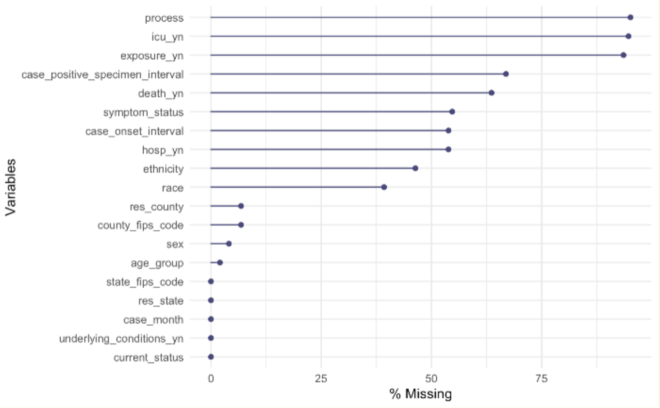


Figure 1. A plot of missing values in each variable. The rows are ordered to show which variables have the most percentage of missing data.

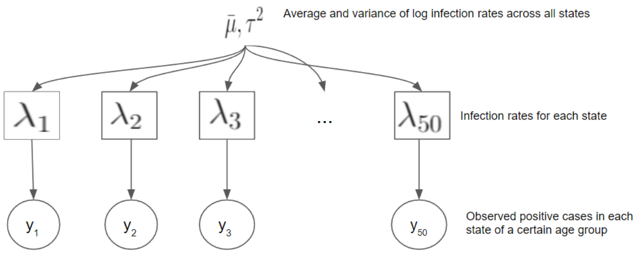

$\bar{\mu}, \tau^2$
Average and variance of log infection rates across all states
λ_1
λ_2
λ_3
...
λ_{50}
Infection rates for each state
y_1
y_2
y_3
y_{50}
Observed positive cases in each state of a certain age group

Figure 2. Hierarchical model for λ_j's

4.1. Model Setting

We decided to use a Poisson distribution to model the infection rate of a certain age group at a state. As shown in the formula below, we use Y_j to represent the number of positive cases reported of a certain age group at state j. We set $n_j\lambda_j$ as the Poisson distribution parameter. The λ_j here is the number of positive cases of a certain age group in state j per 100,000 people. The n_j here is used to standardize the number of positive cases. We choose to add n_j here because we found that the number of positive cases is quite large and varies among different states due to the population difference among the states.

$$Y_j|\lambda_j \sim Poisson(n_j\lambda_j)$$

To achieve the goal of estimating infection rates in each state, we decide to use partial polling which would help to stabilize the estimation and avoid the influence caused by the population difference among states. A hierarchical log-normal model of λ_j's is fitted to perform the partial polling.

In Figure 2, we can see the parameters of this hierarchical model of λ_j's. At the top level, we have the mean $\bar{\mu}$ and variance τ_2 of the $log(\lambda)$s of each state. Since the infection rate in our case is always positive ($\lambda > 0$), it is more reasonable for us to use $log(\lambda)$ with this normal distribution. On the second level, we have the infection rates λ_j of each state. On the third level, there is the raw data of the number of positive cases of a certain age group in each state.

$$\bar{\mu} \sim N(\mu_0, \sigma_\mu^2)$$
$$\tau \sim half - t_7(A)$$
$$log(\lambda_j)|\bar{\mu}, \tau \sim N(\bar{\mu}, \tau^2) forj = 1, ..., 50$$
$$Y_j|\lambda_j \sim Poisson(n_j\lambda_j) forj = 1, ..., 50$$

After determining the parameters of the hierarchical model of λ_j's, we construct our full model with the above formulas presented. First of all, We decided to use a normal distribution with mean μ_0 and variance σ_μ^2 to model $\bar{\mu}$, the mean of $\log(\lambda)$'s. As for τ, the variance of the $\log(\lambda)$'s, a half-t distribution with degree of freedom 7 and parameter A. Here we default the degree of freedom to 7 based on empirical experience. For each of the $\log(\lambda)$'s, we use the normal distribution with mean $\bar{\mu}$ and τ^2 to model them. And for the Y_j's, a Poisson distribution with parameter $n_j\lambda_j$ is applied.

For our model, we need to tune the values of μ_0, σ_{mu}, and A as the prior hyperparameters to match the distribution of our raw data. Then we want to compute the posterior of $(\bar{\mu}, \tau, \lambda_1, ..., \lambda_{50})|Y_1, ..., Y_50$ to perform our estimation on the infection rate of certain age group at a certain state.

4.2. Model 1

For out first model, we fitted with the data from age_group 65+ and we first apply the prior predictive tuning to fit a reasonable prior model. We tune four hyperparameters here: mu0, sigma_mu, nu, and A.

As we can see from above graphs. For setting 1, we use [mu0=0, sigma_mu = 0.5, nu = 7, and A=0.1]. Since the mean of the raw_rate is 1.675659, this prior places around 14.55 % prior probability on a raw rate exceed the mean. Also, since the mean for this model is 1.14, this first prior specification may not be super reasonable.

For setting 2, we use [mu0=0.5, sigma_mu = 0.1, nu = 5, and A=0.1]. Since the mean of the raw_rate

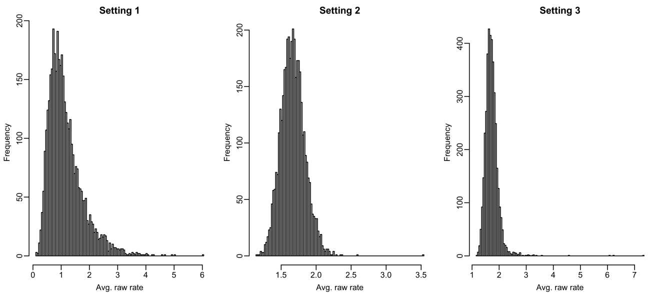


Figure 3. A plot to show three different sets of tuning hyperparameters.

is 1.675659, this prior places around 47.7 % prior probability on a raw rate exceed the mean. Since the mean for this model is 1.673551, which is pretty close to the actual mean. We might think this set of tuning is successful. However, we still want to have a larger span of the x axis which is the prior average raw rate. In that case, we increase A to 0.2 for setting3 to get a wider x span. After all of that, we finished our prior model tuning.

Then, we try to fit the model first using the naive implementation with code in stan file. We using our third specification as the data. As we can see from figure 4 below, the paired plot shows that there were no clear divergent transitions shows in the plot. If there was, there will be red dots which indicate stan has issues proposing next point.

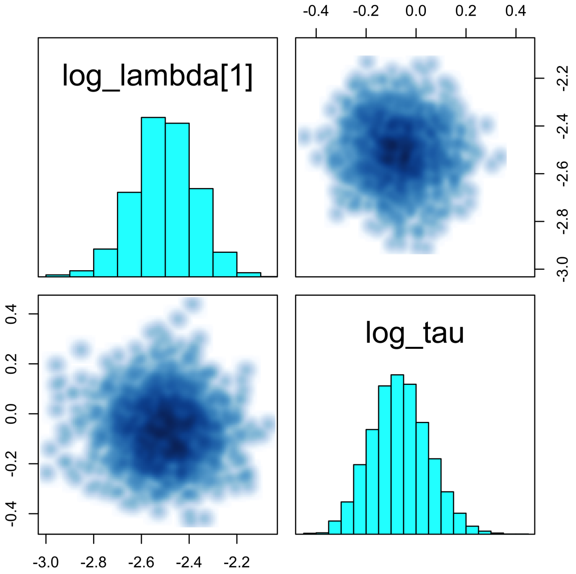


Figure 4. A plot to show if there are divergent transition exist.

After that, we did the prior predictive check for our model. As we can see from the figure below, each row corresponds to simulating a ”fake” set of adjusted case number in each state using our fitted model. The step for prior predictive check is: 1, Draw $log(\lambda)$ from $N(\mu_0, \tau^2)$. 2, compute λ from $log(\lambda)$. 3, Draw simulated dataset Y from Poisson distribution And the below graph shows the histogram of covid cases of people age 65+ in several states of U.S.. Since the orange lines-the original Y value-are all in the bulk, we can say that our simulated dataset resemble the observed dataset.

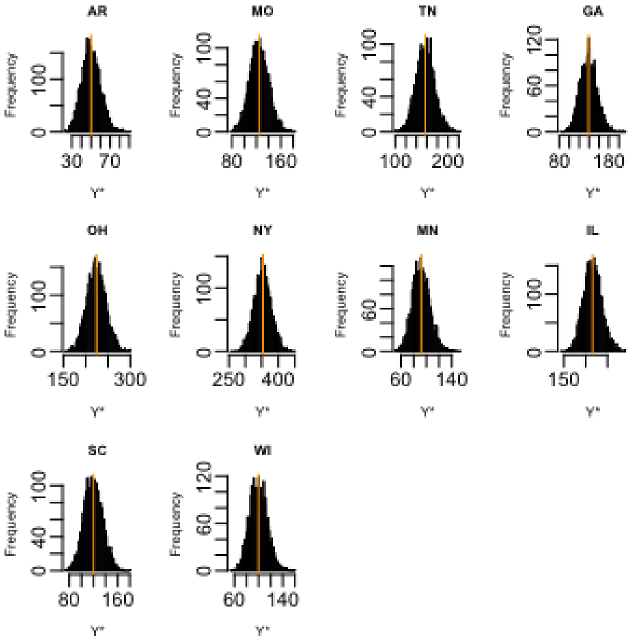


Figure 5. Prior Predictive Check.

After we check our prior predictive model, we can perform the posterior boxplot for this specific age group. We order the state on the x axis in the ascending manner as we can see from figure 6. And it’s quite interesting to see that state Misouri has the highest raw rate.

4.3. Model 2

Pretty similar to our first model, we want to fit a model using data from age group 0-17. We tuning the hyperparameters to [mu0=0.5, sigma_mu = 0.5, nu = 5, and A=0.1] after several adjustment as it shows in the graph below in figure 7. Then we conduct the similar naive implementationa and draw the pairs graph to see if there are any potential divergence. Luckily, there are no red dot shows in the graph and

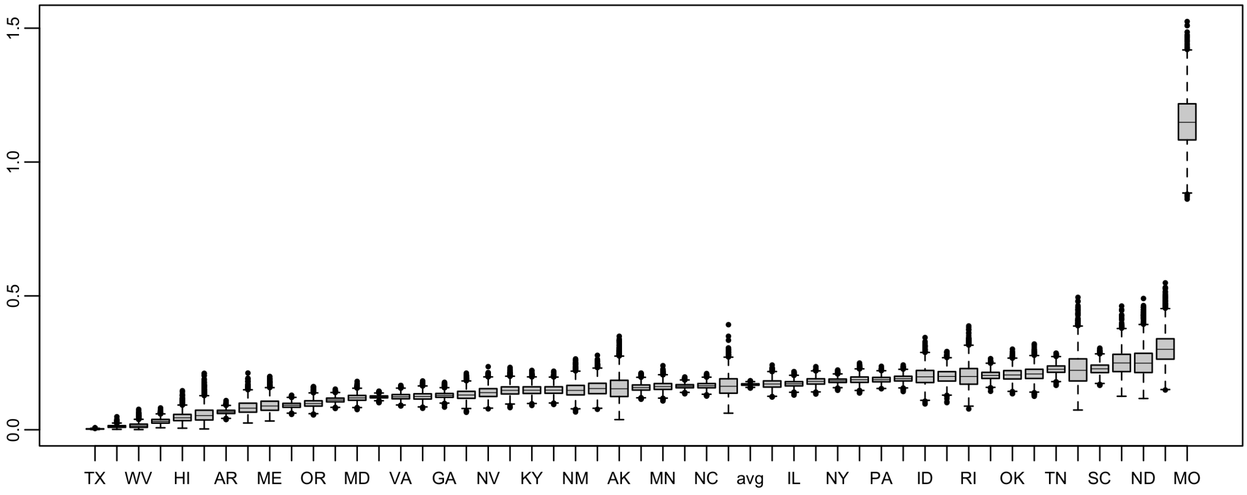


Figure 6. posterior boxplot for 65+.

the prior predictive check is also pretty decent.

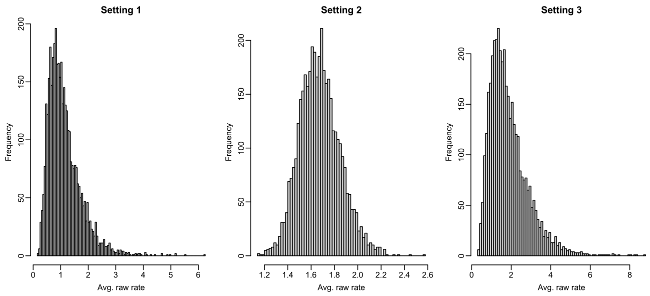


Figure 7. A plot to show three different sets of tuning hyperparameters for model 2.

5. Cross Model Comparison

Then, we have out posterior boxplot for age group 0-17 and we can compare it with our previous one. And as we can see from figure 8 and 9, There are some state remain the same rank in two age groups and some states change their rank. One thing which is pretty obvioud is that for both age group, State Misouri always has the highest case rate. We wonder if our model is actually correct in the real situation. So we search on the internet to see what happened in Misouri, and the below news in figure 10 shows up. Which can prove that our model is good enough to do some research.

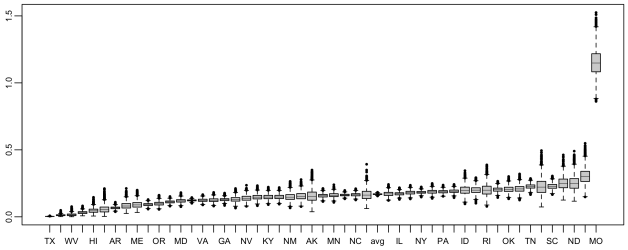


Figure 8. posterior boxplot for 65+.

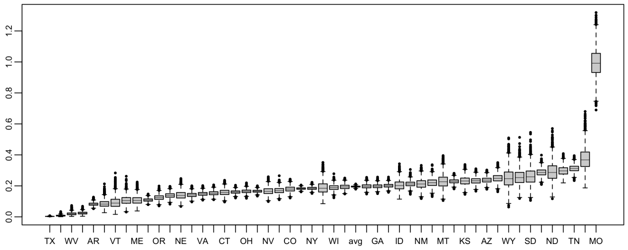
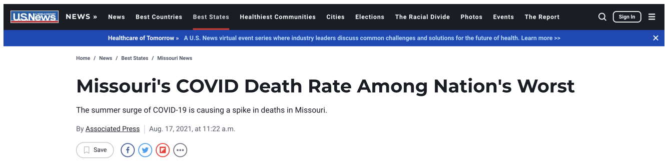


Figure 9. posterior boxplot for 0-17.



Healthcare of Tomorrow » A U.S. News virtual event series where industry leaders discuss common challenges and solutions for the future of health. Learn more >>

Home / News / Best States / Missouri News

Missouri's COVID Death Rate Among Nation's Worst

The summer surge of COVID-19 is causing a spike in deaths in Missouri.

By Associated Press | Aug. 17, 2021, at 11:22 a.m.

Figure 10. US News Report.

6. Comparison between states

In Figure11 and figure12, we make comparison between Wisconsin and New York state by presenting the difference in infection rate for age group $65+$ and age group $0-17$. For age group 65+, the infection rate of New York state is greater than Wisconsin in 71.525 percent of samples.For age group 0-17, Wisconsin's infection rate is greater than New York state in 59.425 percent of samples. After comparing the actual rate for age group $65+$, we find the real data follows the same conclusion in our posterior model.

6.1. Policy Analysis

To find out whether the difference in infections rate between these states are caused by policy differences, we collect all the information of COVID related policy in Wisconsin and New York state, including COVID-19 Vaccine Mandates, School-Related COVID-19 Policies, Social Distancing Actions,

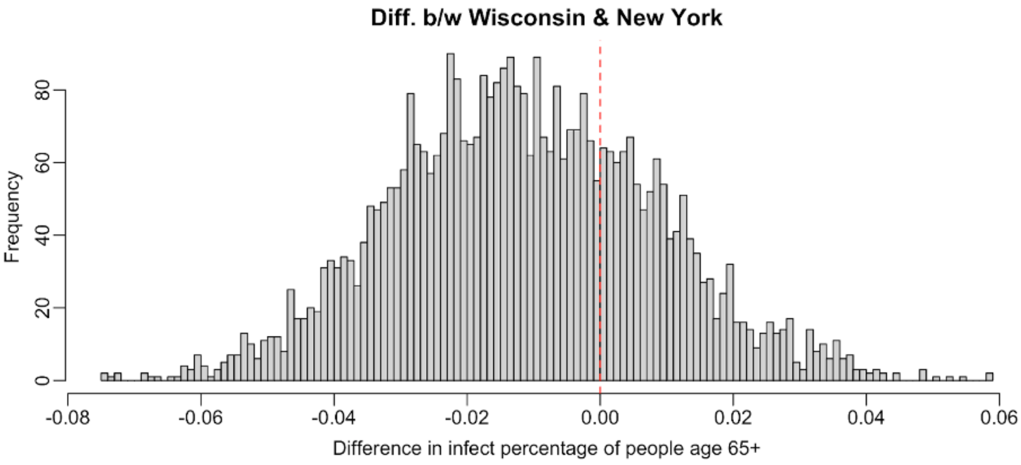


Figure 11. Difference in infection rate of age 65+

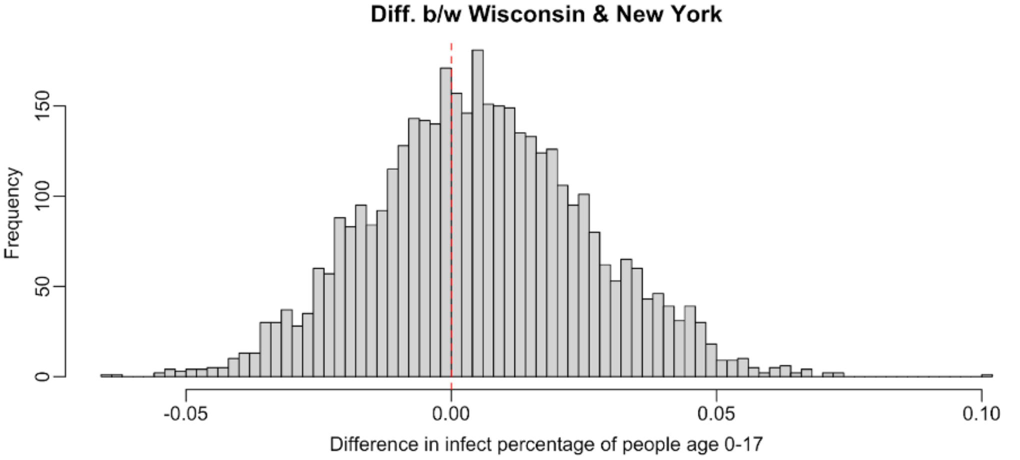


Figure 12. Difference in infection rate of age 0-17

COVID-19 Health Policy Actions, and additional State-Level Data. Here we focus on the difference in infection rate for age group $65+$. We do find differences between policies in Wisconsin and New York state that may affect the infection rate. And differences are mainly found from the following two aspects:

- COVID-19 Vaccine Mandates in place: Wisconsin currently have no vaccine mandate. For New York State, all state-employee must submit proof of COVID-19 vaccination or undergo weekly COVID-19 testing.

- Social distancing Actions: No statewide face mask requirement for Wisconsin. But some counties in Wisconsin do have indoor mask mandates. For New York State, there do exist statewide face

State	Vaccination rate
Wisconsin	99.3 %
New York State	91.7 %

Table 1. Difference in Vaccination rate

mask requirements for people who are not vaccinated.

Although differences do exist in policies in Wisconsin and New York state, it is not strong enough to support our finding as the infection rate of people older than 65 in Wisconsin state is lower than New York state. That is because New York State even enforces more strict rules than Wisconsin in Vaccination mandate and social distancing actions in most aspects. Vaccination was supposed to have an effect on the infection rate. But one interesting finding about vaccine policy is, both Wisconsin and New York state are one of several states which provides free cost vaccination when available. So, there may exist other factors which contribute to the difference in infection rate for people older than 65. We will further explore these factors.

6.2. Possible Explanation

After collecting Covid-19 data between Wisconsin and New York state, we find there is a great difference in the vaccination rate for age $65+$ between these states. According to the Mayo clinic[2], the vaccination rate of people older than 65 in Wisconsin is 99.3 percent, compared to the vaccination rate of elder people in New York state, which is 91.7 percent. This difference in vaccination rate do influence the risk of being infected for elder people. In the study conducted by Moghadas et al. in 2020, they evaluated the impact of a 2-dose COVID-19 vaccination on reducing infection and other risks. And they found the the highest reduction on attack rate happened among people older than 65, which is 54 to 62 percent[4]. Therefore, vaccination do have a great improve in preventing elders being infected or facing other risks. At the same time, Wisconsin's lower infection rate for age group $65+$ can be caused by the its higher vaccination rate.

In addition to explain the difference in infection rate, we further find evidence to explain the the great difference in vaccination rate, which is poverty rate. The updated data posted by America's health ranking[1] shows that the poverty rates for people older than 65 in Wisconsin and New York state are 7.4 percent and 12 percent. The average poverty rate in the US is 9.4 percent. So, Wisconsin has a much lower poverty rate for older people than New York state. According to Freed et al.(2021), vaccination rates among adults ages 65 and older are 10 percent lower in counties where a relatively high proportion of adults 65 and older live in poverty than counties where a lower share of older adults live in poverty[3]. Therefore, the high poverty rate of elder people will negatively affect their vaccination rate. The great difference in poverty rate between these two states explain the gap between their vaccination rates, which further be related to their infection rate.

7. Conclusions

In conclusion, our hierarchical model analysis allows us to compare on the Covid-19 infection rate between different age groups in different states. We can draw some general conclusions on both aspects from the current analysis. For example, in terms of age groups, age group 65+ has lower case rate compare to age group 0-17, and in terms of the states, Missouri has the largest case rate among others for both age groups, 0-17 and 65+.

7.1. Limitation and Future Direction

There are some limitations existing in our project. First, we only conducted the analysis on age group 0-17 and 65+. Further developing models on other age groups is worthwhile. Through comparing infection rates of more age groups, we can have a more comprehensive analysis about one state, such as people from which age group are most easily infected in a specific state. Also, there will be more evidence supporting us to conduct a more diverse policy analysis such as whether some policies has significant effects on one specific age group. Besides, we found that not only policies are decisive on a state' s infection rate but other factors such visitors' flow rate of each state also. More research and analysis can be conducted based on our model and results.

References

[1] Explore poverty - ages 65+ in wisconsin — 2021 senior report ... 2021.

[2] U.s. covid-19 vaccine tracker: See your state's progress. 2021.

[3] J. C. J. T. T. N. Meredith Freed, Kendal Orgera. Vaccination rates are relatively high for older adults, but lag in counties in the south, in counties with higher poverty rates and in counties that voted for trump. 2021.

[4] K. Z. C. R. W. A. S. B. H. S. L. A. M. K. M. N. J. M. L. M. C. F. A. P. G. Seyed M. Moghadas, Thomas N. Vilches. The impact of vaccination on covid-19 outbreaks in the united states. 2020.