# Final Report

Yixuan Wang, Tongyue Jia, Qingjie You, Siddhant Thakur

April 2021

## 1 Background

From the beginning of 2020, a new epidemic virus named COVID-19 has swept the world and influenced our daily life. Our school, the University of Wisconsin-Madison, has been closed for almost a year. To reopen the campus safely as soon as possible and return to our campus life before the epidemic, i is essential to give an effective and suitable strategy for vaccinating the university population and the Dane County area.

The key points to consider while reopening our campus are safety and time. We want to ensure that everyone, both on-campus and off-campus, takes vaccinations as quickly as possible. In the meanwhile, we want to have fewer people infected during the arrangement of vaccinations.

## 2 Strategy

First, we will explain the goal of the whole vaccinating process. To achieve this goal, we need to allocate vaccines to different area and then within each area, we decide the vaccinating order among different groups of people. Our strategy will focus on this two major problems in vaccinating process.

**1. Goal: The goal of the whole vaccinating process is to achieve a 100% vaccinating rate on campus and a 70% vaccinating rate outside the campus.** According to, 71% of American people are willing to take vaccines[1]. The CDC's research shows that a 60% vaccinating rate is needed to achieve herd immunity[2]. So we take 70% as the expected vaccinating rate outside the university to ensure that the surrounding environment in Dane County is safe enough for the university to reopen.

**2. Vaccines allocation:**

- **Vaccines are dynamically allocated to each municipality every week, and we assume that these vaccines are injected within one week.** We decide the amount of vaccine required after analyzing the data and setting our vaccination priorities.
- **We only consider the dynamic allocation of the first dose.** For the second dose, we allocate the vaccines according to the distribution of the first dose.
- **Vaccines allocation inside the university:** We allocate a fixed number of vaccines to the university every week until achieving a 100% vaccinating rate on campus. The allocation of vaccines inside the university should consider the maximum capacity of vaccines injected in the school. From the dashboard about the weekly doses administered by UHS[3], we can find that in the recent four months, the maximum injection by UHS is 2729. Thus, we can allocate 2700 as the fixed number of vaccines to the university until the 100% vaccinating rating on campus.
- **Vaccines allocation outside the university:** For allocating the first dose vaccine every week, we determine the planned injection rate proportionate to the existing positive rate on the last day of the last week in each municipality. We assume that the newly increased infection rate is positively correlated to the existing positive rate. Notice that we consider rate instead of the total number to exclude the influence of the population number in each municipality. When the existing positive rate rises, the probability of contact with the infected people and their close contacts will also rise. So the newly increased positive rate tends to rise in the area. In order to prevent this trend, more vaccines should be allocated to this area. Therefore the planned vaccinating rate should be proportionate to the existing positive rate. When the vaccinating rate has achieved 70% in a municipality, we will not consider this municipality when allocating future vaccines.

**3. Vaccinating order:**

After allocating vaccines to the university and different municipalities outside the campus, we need to determine the vaccinating order among different groups of people. We divide all the population(including professors, staff, students) into different age groups, then we vaccinate different age groups in order. We design this order according to the positive rate of different age groups. We will examine whether age is an influencing factor in the likelihood of infection. If it is so, we will vaccinate the most susceptible age group first, then other groups.

# 3 Question Refinement

The core of this strategy based on two assumptions:

- The newly increased infection rate is positively correlated to the existing positive rate.
- Age is an influencing factor to the likelihood of infection.

If we can validate these two assumptions by data analysis, then we confirm this strategy's effectiveness. So we refine the best strategy problem to two specific questions:

- Is the newly increased positive rate positively correlated to the existing positive rate in Dane County?
- Is age an influencing factor to the cumulative positive rate among people in Wisconsin?

In the latter part of this report, we will describe our data sets models to examine these two refined questions.

# 4 Data

## 4.1 Data for Question1

We want to test whether there is a positive correlation between the existing positive rate and the newly positive rate. To ensure that our conclusion is instructive for vaccine allocation at the present stage, we should use the existing positive rate and newly positive rate in the past 30 days in Dane County.

The initial data set contains the number of cumulative positive cases, the number of newly positive cases, and the number of cumulative deaths. To prove a positive correlation between the existing positive rate and the newly increased positive rate, we need to calculate these two variables based on our initial data. Firstly, we calculate the number of existing positive cases:

$$\begin{aligned} Existing\ Positive = {} & Last\ Day\ Existing\ Positive + Newly\ Positive \\ & - Newly\ Deaths - Newly\ Recovery \end{aligned} \tag{1}$$

However, we are unable to calculate this recursive formula since the "Newly Recovery" data is unavailable. The "Newly Recovery" data is hard to collect accurately since some people just stayed at home to recover and the sign of recovery is still not generally acknowledged. According to DHS[4], the recovery cases are defined as the number of confirmed cases that are currently alive based on Wisconsin state vital records system data and have 30 days since symptom onset or diagnosis. So we estimate the number of the existing positive cases as below:

$$\begin{aligned} Existing\ Positive = {} & the\ Sum\ of\ Newly\ Positive\ Cases\ in\ Last\ 29\ Days \\ & - the\ Sum\ of\ Deaths\ in\ Newly\ Positive\ Cases\ in\ Last\ 29\ Days \end{aligned} \tag{2}$$

There is a problem in tracing the test date of each death case from the original data, so the sum of deaths in newly positive cases in the last 29 days is unattainable. Since we only use the data collected in the last two months and the death cases are very few in this period, we ignore the minor differences between the sum of deaths in newly positive cases in the last 29 days and the sum of newly deaths in the last 29 days. Therefore, we draw the approximate formula for the number of the existing positive cases:

$$\begin{aligned} Existing\ Positive &= the\ Sum\ of\ Newly\ Positive\ Cases\ in\ Last\ 29\ Days \\ &- the\ Sum\ of\ Newly\ Deaths\ in\ Last\ 29\ Days \end{aligned} \tag{3}$$

Finally, we need to get the existing positive rate and newly positive rate based on dividing the number of existing positive and the number of newly positive by the total population in Dane County, and the total population estimation in Dane county is 546,695 on the U.S. Census Bureau[5]:

$$\begin{aligned} Existing\ Positive\ Rate &= \frac{Existing\ Positive}{Total\ Population} \\ Newly\ Positive\ Rate &= \frac{Newly\ Positive}{Total\ Population} \end{aligned} \tag{4}$$

And there is another definition about newly positive rate which is called 7-days average newly positive rate. The number of 7-days average newly positive is defined as

$$7-days\ Average\ Newly\ Positive_t = \frac{1}{7}\sum_{i=1}^{7} Newly\ Positive_{t-7+1} \tag{5}$$

and the rate is defined as same as equation (4) above.

Because the scale of these two rates is so small that it cannot be well expressed and may lose some accuracy, we use another expression of these two rates in our model with the equation $Rate = rate * 1000$. This new rate represents how may people are the existing positive or newly positive out of every thousand people. Thus, from all these steps, we can get the data set which will be used to build our model. This data set includes the existing positive rate and newly positive rate in the last 30 days of records (Feb $21^{th}$ 2021 to Mar $21^{th}$ 2021) in Dane County.

## 4.2 Data for Question2

For this question, we want to determine whether different age groups could divide people into specific order of their susceptibility to the virus. According to the data on the DHS website, we have data sets in different counties in Wisconsin. The data set included the cumulative total positives cases and cumulative positive cases by age groups in each county with an interval of 10. Because the discrepancy existing in the total cumulative positive cases of different counties would interfere with our testing results, we use

$$Positive\ cases\ Ratio = \frac{Group\ positive\ cases}{Total\ positive\ cases} \tag{6}$$

to represent the discrepancy among positive cases under each age group in each county instead of just using the number of positive cases. The univariate plot for each age group against each county's number of positive cases is shown in the following figure.

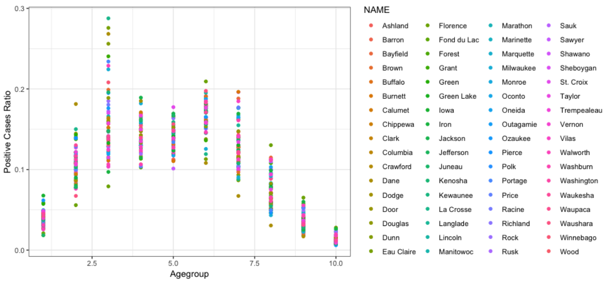


Figure 1: Multivariate plot for each age group in each county

# 5 Model

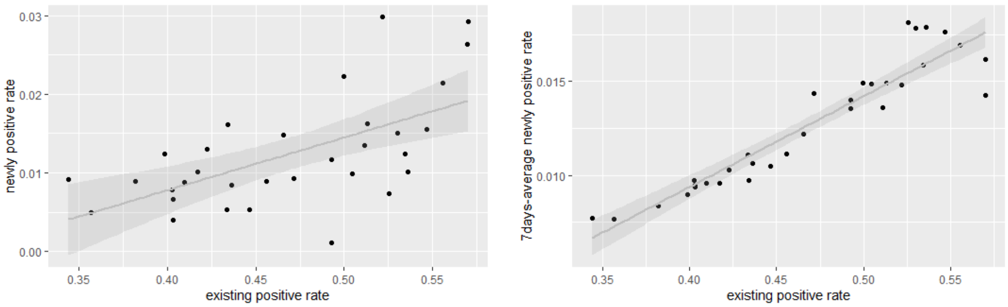
## 5.1 Model for Question1

Firstly, we consider correlations between the newly positive rate day by day (called newly positive rate for short) and the existing positive rate (Model 1). At the first step, we check whether all the data sets are normal and then decide which method we should use in calculating the correlation coefficients. After confirming that all these data sets we are using come from a normal distribution, we use the Pearson method to calculate correlation coefficients. Pearson correlation coefficient is a statistic used to describe the degree of correlation between two variables, and it is applied on continuous variables following a normal distribution. It will return a symmetric correlation matrix with all diagonal elements equal to 1. The correlation matrix is as Table 1:

| | Existing Positive Rate | Newly Positive Rate |
|---|---|---|
| Existing Positive Rate | 1 | 0.6065165 |
| Newly Positive Rate | 0.6065165 | 1 |

Table 1: Model1: Correlation Matrix from Pearson Test

The estimation of the correlation coefficient for this model is positive. Therefore, we can imply a positive correlation between the existing positive rate and the newly positive rate. We can use the data visualization to show its correlations (Figure 2(a)).



(a) Model 1

(b) Model 2

Figure 2: Correlation Plot

The straight line in Figure 2 is the fitted line of the linear regression model. It shows that when the existing positive rate increases, the newly positive rate will also increase. Other than that, the shadow part is the 95% confidence interval of the linear regression. So, the correlation coefficient and the plot of this model both prove the positive correlations between the existing positive rate and the newly positive rate, which is the evidence of our first refined question. The proof for refine question 1 shows that it is reasonable to distribute our vaccines based on the existing positive rate as we mentioned in **Question Refinement** part.

However, we find that nearly half of the points are outside the confidence interval. This means that the newly positive rate predicted by the linear regression is quite different from the actual data. We are trying to find the reason behind this difference.

After re-observing the newly positive data, we found a periodic phenomenon for the number of newly positive cases. The period time is one week which is showed in Figure 3[6], one day with a large number at the beginning, and two days with a smaller number relatively in a cycle. Combining with the date in the data set, we found that the number of newly positive is huge every Monday and smaller on the weekends.

Based on this periodic phenomenon, we decided to use 7-days average newly positive rate to replace the newly positive rate day by day in order to reduce the influence from the periodic phenomenon, which is called Model 2. We repeated the step above and got the following correlation coefficients matrix (Table 2), and the data visualization is in Figure 2(b).

| | Existing Positive Rate | 7-days Average Newly Positive Rate |
|---|---|---|
| Existing Positive Rate | 1 | 0.9370116 |
| 7-days Average Newly Positive Rate | 0.9370116 | 1 |

Table 2: Model2: Correlation Matrix from Pearson Test

The correlation coefficient for model 2 is larger than for model 1, and there are more points out of the 95% confidence intervals for Figure 2(a) than Figure 2(b). This phenomenon shows that model 2 may be better than model 1.

These two models both prove the positive correlations between existing positive rate and newly positive rate, which is equivalent to that highly existing positive rate will lead to highly newly positive rate. This result proves our refined question 1, and gives an credible reason to distribute the vaccines based on the existing positive rate in order to hinder the increasing of newly positive. However, this better trend in model 2 may be related to the fact that we used 7-days smoothing (”7-days Average Newly Positive Rate”) as the observation to eliminate the impact of periodicity. For example, during the weekend, the decrease in the number of newly positive cases is probably due to the decrease in the number of tests on non-working days, rather than the decrease in the existing positive rate. Using a 7-days average, we can effectively take care of this event.

## 5.2 Model for Question2

For this question, we want to determine whether different age groups could divide people into specific order of their susceptibility to the virus. After checking the data type we get in Part3.2, we can group data by age groups, which means we can use the multilevel model to estimate different age groups' impact on the positive cases. The model can be formed as

$$\text{logit}(Y_{ij}) = \mu_j + \epsilon_{ij} \qquad \mu_j \sim N(\mu, \sigma^2) \qquad j = 1, ..., n \tag{7}$$

$Y_{ij}$ denotes the $i$th positive cases ratio, which resides in the $j$th group. $n$ is the total number of groups. However, to determine the vaccination order, especially in the university, we do not need such detailed classifications with interval of 10. Therefore, we also apply our models to the age group with an interval of 20. The results are shown in Table 3 and Table 4.

| | Intercept | | Intercept |
|---|---|---|---|
| Age0-9 | 0.03980473 | Age50-59 | 0.16611479 |
| Age10-19 | 0.10757445 | Age60-69 | 0.12935918 |
| Age20-29 | 0.15581286 | Age70-79 | 0.07239797 |
| Age30-39 | 0.14138382 | Age80-89 | 0.03527745 |
| Age40-49 | 0.13954784 | Age90+ | 0.01272691 |

Table 3: Age Group with each interval = 10

| | Intercept |
|---|---|
| Age0-19 | 0.14759283 |
| Age20-39 | 0.29776770 |
| Age40-59 | 0.30598842 |
| Age60-79 | 0.20132108 |
| Age80+ | 0.04703585 |

Table 4: Age Group with each interval = 20

Since higher value of intercept means the corresponding age group has a higher positive cases ratio, based on the results in table 3, people in age50-59 will have the highest priority to take vaccinations. In contrast, people older than 90 will have the lowest priority to be vaccinated. However, for a broad classification in Table 4, we can see that the order of vaccination for some people has changed. For example, in Table 3, people in age40-49 are in the forth place and have risen to the first place in Table 4. We can use Simpson's Paradox to explain this phenomenon. When we shrink the age40-49 and age50-59 into one group, age40-59, the effect of age40-49 is less than the effect of age50-59. That is to say, the final probability will almost depend on the probability of age50-59 which ranks first in Table 3.

Besides, from the figure 1, we can know that the positives cases ratios also vary on different counties. That is to say, we should consider different counties as factors which could also impact the positive ratios for an age group. Thus, we introduce the multinomial model to address this problem.

In this case,the positive cases ratio could be viewed as the probability of a positive case being classified into an age group from a specified county. According to the background, we want to know the probability of each age group in Dane county. Under this situation, we find that the outcomes follow a multinomial distribution and the model can be formed as:

$$P(\text{majority is } k|\text{observation } i] = \begin{cases} P_{ki} & k \in [Age10-19, ..., Age90+] \\ 1 - \sum_{k\in[Age10-19,...,Age90+]} P_{ki} & k = Age0-9 \end{cases} \tag{8}$$

For each $P_{ki}$ we can define $\tilde{g}_k(P_{ki}) = X_i'\beta$. $X_i$s are intercepts and counties. Thus, we can use multinomial regression to determine the probability of each age group in Dane county, shown in Table 5. To compare the results from the multilevel model, we also apply our models to the age group with an interval of 20, shown in Table 6.

| | Probability | | Probability |
|---|---|---|---|
| Age0-9 | 0.04000120 | Age50-59 | 0.15999836 |
| Age10-19 | 0.11000074 | Age60-69 | 0.12999941 |
| Age20-29 | 0.13999935 | Age70-79 | 0.07999887 |
| Age30-39 | 0.15000161 | Age80-89 | 0.02999870 |
| Age40-49 | 0.15000116 | Age90+ | 0.01000059 |

Table 5: Age Group with each interval = 10

| | Probability |
|---|---|
| Age0-19 | 0.15000097 |
| Age20-39 | 0.29000023 |
| Age40-59 | 0.30999813 |
| Age60-79 | 0.21000025 |
| Age80+ | 0.04000042 |

Table 6: Age Group with each interval = 20

If people are grouped by an interval of 20 in the multinomial model, the value of the estimated intercepts for each group in Table 6 shows that the order of each group to take vaccinations is the same as the order we conclude from the previous multilevel model. However, if people are grouped by an interval of 10 in the multilevel model, the priority of vaccination is different from what we get in the multilevel model. For example, age20-29 has a lower priority than the age30-39 and age40-49 in this model. Simpson's Paradox could also be one of the reasons to explain such a difference in the result. Another point is that we incorporated county as a predictor in the multinomial model. However, in the multilevel model, different counties are not considered to affect the positive cases ratio of an age group.

Both the two models are using the cumulative positives cases, which means we didn't consider the number deaths as an factor to determine the order of vaccinations. However, for people who is greater than 80, they are more likely to die than other infected people[5]. Hence, one of the limitations of our models is that we couldn't consider both positive cases and deaths to get an vaccination order for age groups. Another limitation is the Simpson's Paradox we used to explain the model results. We didn't get an specified value to determine such weight.

# 6 Conclusion

For the first refined question, we confirmed that there is a positive correlation between existing positive rate and the newly positive rate. Based on what we have mentioned in strategy part (Outside the university), we assume that this positive correlation is the basic principle that affect the vaccine distribution. Thus, we plan to distribute the vaccines based on the existing positive rate for each area in Dane County, which may hinder the increasing of the newly positive numbers in high-risk areas.

For the second question, we can know that different age groups will have different effects on the positive cases rate and we can determine the order of vaccinations based on different age groups. The priority of taking vaccination from the highest to the lowest is that people in age50-59 first, then age30-39, then age40-49, then age30-39, then age20-29, then age60-69, then age10-19, then age70-70, then age0-9, then age80-89 and finally age greater than 90. Especially in university, we do not need such detailed classifications because most of people are in age 20-59. Thus, we can just follow the results in Table6. People in age40-59 rank first, age20-39 rank second, age60-79 rank third, age0-19 rank forth and age80+ rank fifth.

# References

[1] Path to Normality : 2021 Outlook of COVID-19 in the US. (2021, March 8), from https://covid19-projections.com/path-to-herd-immunity/

[2] Americans' willingness to get COVID-19 vaccines reaches record high: Poll. (2021, February 11), from https://abcnews.go.com/Health/americans-willingness-covid-19-vaccines-reaches-record-high/story?id=75807568

[3] The weekly doses administered by UHS: https://covidresponse.wisc.edu/dashboard/

[4] The definition of recovery on DHS: https://www.dhs.wisconsin.gov/covid-19/cases.htm#recovery

[5] The total population estimate in Dane County: https://www.census.gov/quickfacts/danecountywisconsin

[6] This plot is used to show that there is a periodicity in the newly positive:

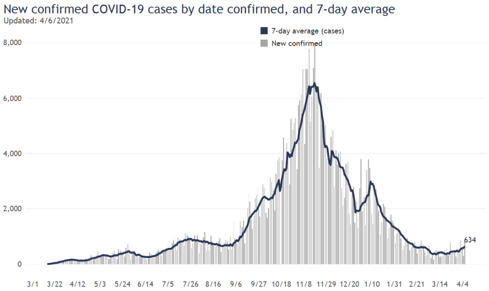


Figure 3: The Number of Newly Positive to show Periodicity