COVID-19의 백신 예방 접종 효과에 대한 데이터 분석

이재욱[1], 함승준[2]
[1]청심국제중고등학교, [2]Worcester Academy

Analyzing the Data behind the Vaccination Efficacy of COVID-19

[1], [2]
[1]CheongShim International Academy, Gyeonggi-do, Korea
[2]Worcester Academy, Massachusetts, USA

요 약 정부의 광범위한 백신 접종 노력 이후 코로나 19 바이러스의 전반적인 확산은 감소했지만, 백신 접종과 바이러스의 확산 추세 사이의 상관관계에 관한 연구는 거의 수행되지 않았다. 이 연구에서는 코로나 19 팬데믹의 전파 상황을 분류하기 위해 각각 확산 추세 (Continued trend), 통제 추세 (Sustained trend), 그리고 평균적인 추세 (Average trend of previous cases)의 세 가지의 시나리오를 제시했다. 이 연구에서는 백신 접종을 받은 개인의 수와 일일 사례 수의 감소 사이의 상관관계가 존재한다고 가정했으며, 연구를 통해 확산 추세에서 백신 접종이 이루어질 때 79.12%로 가장 높은 감소율을 보였고 평균 추세에서 감소율은 63.66%로 가장 낮았다는 것을 밝힐 수 있었다. 이 연구를 통해 백신 접종과 COVID-19 감염 사이에 존재하는 명확한 상관관계를 밝히고, 백신 접종의 확대를 뒷받침할 수 있는 근거를 제공하고자 한다.

Abstract Although the overall spread of the COVID-19 virus has declined since the extensive vaccinations, little research has been done on the correlation between vaccination and the spread of the virus. In this study, to classify the spread of the COVID-19 pandemic, three scenarios were presented; continuous trend, sustained trend, and average trend of previous cases. In this study, it was assumed that a correlation existed between the number of vaccinated individuals and decreased number of daily cases. This study showed that the highest reduction rate was 79.12% when vaccination occurred in a continuous trend, and lowest reduction rate was 63.66% in an average trend. This study aims to elucidate the clear correlation that exists between vaccination and COVID-19 infection and to provide evidence to support the expansion of vaccination.

Key Words: COVID-19, Vaccination, Population Scale, Data Analysis

1. Introduction

1.1 Characteristics of Online Hate Speech

Ever since the emergence of the recent Coronavirus, public facilities have been shut down in order to enforce strict quarantine and countless mortalities have resulted from the rapid spread of the virus. The detrimental effects of this pandemic have not been localized to one certain region, but rather have had a vast global effect since its appearance in November 2019. In the early stages of the pandemic, scientists from all over the world delved into researching the virus to try to better understand the spread pattern and develop a vaccine, but it was not until four months upon its discovery that the COVID-19 virus had an expedited FDA approved vaccine. Due to the shortage of these vaccines, countries prioritized vaccinating essential workers and high-risk individuals, and the global vaccination process is still in process during the writing of this paper.

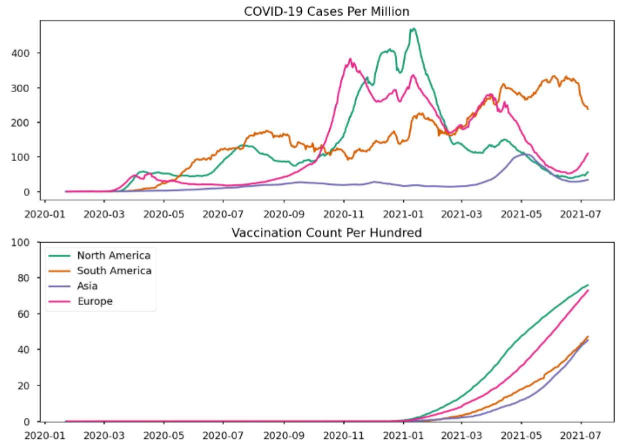
Historically, vaccines were developed to reduce the fatal effects of past epidemics and pandemics as one of the most efficient methods to protect people from infection. A majority of the vaccines (Pfizer, Moderna) require two doses set approximately 2-3 weeks apart in order to maximize immunity build-up. [1] However, with all vaccines, there is no guarantee that the risk of infection is zero percent as many factors come into play to affect the probability, such as exposure to infected individuals and the number of people wearing masks. Herd immunity will take effect once a majority of the population is vaccinated, thus decreasing the number of and interactions with potentially infectious individuals. With a particularly strong mutation of the virus, there is a chance at the appearance of breakthrough infections, a phenomenon in which some vaccinated individuals will become infected with a variation of the virus

The vetting process for vaccinations requires at least a 50 percent efficacy in order to be approved. The recently passed COVID-19 vaccines range from 50.38% to 100% (depending on the disease severity or trial) [2] and did not go through the same approval process as other vaccines; the requirements for the vaccine approval were lowered given the pandemic nature of the issue. The traditional vaccine development pathway usually takes 10-15 years, but because of the fast spreading pandemic, scientists turned to a faster development pathway that takes 12-18 months. This is done by combining some of the phases and processing the clinical trials faster while keeping the vaccines' safety and accurate data in check. Most of the commonly used vaccines have been approved for Emergency Use Authorization by the World Health Organization (WHO) but have not been approved by the U.S. Food Drug Administration (FDA) with the exception of Pfizer, which is now FDA approved. [3] Efficacy measures the decrease in risk of infection and is calculated through many clinical trials with controlled and placebo groups, first tested in mice before moving on to human trials. Studies conducted by Yale Medicine have shown that the COVID-19 vaccines affect people equally despite the variables of age, race, ethnicity, gender, BMI, among others. Each vaccine type has different efficacy rates but does one common thing in common: it protects the individual from infection.

As mentioned prior, variants of the virus have developed as mutations, further worsening the severity of the pandemic due to the mutations' higher transmissibility and stronger symptoms. Due to the game of catch-up that researchers are having to play to learn about these newly developed strains, the true efficacy of the vaccines in individuals has yet to be researched in depth.

The COVID-19 vaccines are all mRNA vaccines that upon injection equip the human body with a blueprint for cells to produce certain proteins. The cells are tasked with creating a similar protein to that of the coronavirus and the immune system then practices an appropriate response to the protein. If the body is infected with the actual protein, the immune system will launch a quick response due to the previous exposure. This approach is different to traditional vaccines, like the ones for chicken pox or measles, in that the conventional method directly injects antigens into the body and the body then produces antibodies to fight off the inoculated pathogen. Scientists anticipate mRNA vaccines to streamline the vaccination process by taking out the bespoke production means of traditional vaccines that need separate purification and testing for different strains of virus. The mRNA vaccine may allow for just small changes in the encoding of the target protein allowing for quicker vaccine development, but this needs further experimentation before it can become the standard.

Previously published studies conducted on the efficacy of COVID-19 vaccine have less real-world applicable results due to the limited number of test subjects as well as the controlled laboratory environment in which the experiments were conducted. Both of these factors test vaccines in favorable conditions where the results come out to be higher than when released to the public. In a study released in September 2021, a group of researchers conducted a study on the BNT162b2 mRNA COVID-19 vaccine through six months. [4] The BNT162b2 vaccine is similar to the aforementioned mRNA vaccines approved and authorized by the Federal Drug Association for use during the pandemic. The study focused on testing the efficacy of the vaccine against laboratory-confirmed COVID-19 vaccine for six months after initial vaccination of 44,165 subjects. While they maintained diversity in test subjects, the subject size is not big enough to give insight into how larger populations will fare from the vaccine. Thus, we set out to study the efficacy of distributed vaccinations at a population level to pave the way for future insight into the trajectory of the virus and its effect on global populations.



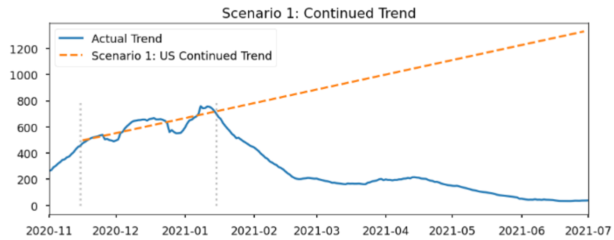
[Fig. 1] Graph of COVID-19 cases per million (top)
Graph of vaccination trends (bottom)

In an attempt to understand the effectiveness of the COVID-19 vaccines better, we collected data on the number of people vaccinated and the number of COVID-19 cases from certain countries within a set time period. Then, we examined the trends of both data sets along with identifying the different factors affecting each country's general trends in spread. From our initial analysis, we found Asia to have a smaller number of daily cases because the information is normalized per population (per million), where the actual counts are the highest in Asia. The information presented in this paper seek to prove whether or not vaccination is the solution to the pandemic and help predict impending spikes in COVID-19 cases in order to better prepare for the situation.

2. Methods

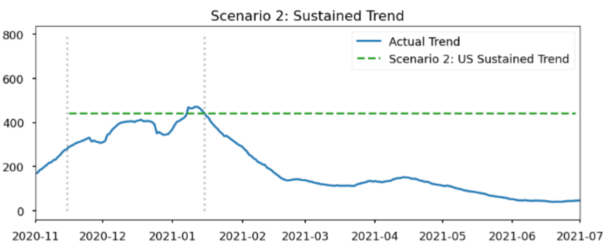
The initial approach to this study was barred by the lack of information available about the trend of unvaccinated COVID-19 cases. The data available at the time of this study was focused on the trend of cases after the mass distribution of vaccines due to the timeframe in which the data was collected. To best navigate through unclear times, we proposed three different scenarios of how the pandemic could proceed: a continued trend, a sustained trend, and the average trend of previous cases. These methods would be used to calculate the actual trend using actual data from a certain time frame in the U.S.

The first scenario was of the Continued Trend. This scenario takes actual data collected from November 15, 2020 to January 15, 2021 and conducts linear regression to calculate the actual trend. The regression coefficient of determination (R2) was 0.67. Though this scenario is most likely an overestimation of the projected trend, this was included to present a worst-case scenario to prepare for all possibilities.



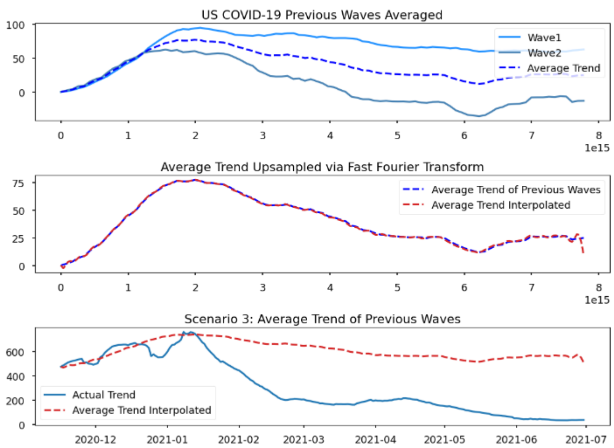
[Fig. 2] Trajectory of Continued Trend in US

The second scenario was of the Sustained Trend. As the time window to infer the post hoc daily COVID-19 cases were from November 15, 2020 to January 15, 2021, this scenario assumes that the daily number of new cases stayed constant beginning from January 15, 2021. From the observed data, this value was determined to be 443 cases per million. The projected trajectory with the value of 443 cases per million in this scenario is optimistic because it works off the idea that the previously upward trend has since flattened.



[Fig. 3] Trajectory of Sustained Trend in US

The final scenario was of the Average Trend of Previous Cases. This scenario calculates the average waveform of the US COVID-19 first wave, dating from March 16, 2020, to June 16, 2020, and the second wave, dating from July 1, 2020, to October 1, 2020. We resampled the average waveform found to match the length of the current waveform of 181 days using scipy.signal.resample which upsamples or interpolates samples in the frequency domain using Fast Fourier Transform. Once the waveform was resampled, it was scaled to match the initial point and maximum point of the actual trend. Of the three scenarios presented in this paper, this is the most ideal as we are unable to know whether the number of cases would have gone down without the introduction of vaccination; the actual trend could have continued to rise before falling in a similar fashion as previous waves without vaccination.

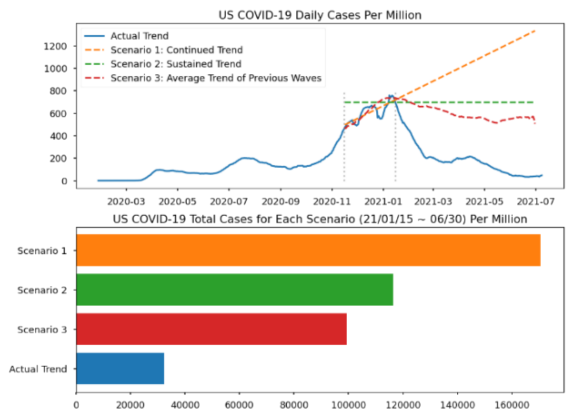


[Fig. 4] Trajectory of Average Trend of Previous Cases

The final scenario was of the Average Trend of Previous Cases. This scenario calculates the average waveform of the US COVID-19 first wave, dating from March 16, 2020, to June 16, 2020, and the second wave, dating from July 1, 2020, to October 1, 2020. We resampled the average waveform found to match the length of the current waveform of 181 days using scipy.signal.resample which upsamples or interpolates samples in the frequency domain using Fast Fourier Transform. Once the waveform was resampled, it was scaled to match the initial point and maximum point of the actual trend. Of the three scenarios presented in this paper, this is the most ideal as we are unable to know whether the number of cases would have gone down without the introduction of vaccination; the actual trend could have continued to rise before falling in a similar fashion as previous waves without vaccination.

3. Results

As seen in Figure 5, the projected trends within our study were significantly higher than the actual trend. In Scenario 1, the Continued Trend, the total cases capped at 170,580 cases per million. Scenario 2, the Sustained Trend, showed a constant in number of cases and resulted in 116,411 per million total number of cases within the designated time frame. Finally, Scenario 3, the Average with 99,427 cases per million. Trend of Previous Waves, showed the smallest difference from the actual trend of our three proposed scenarios. It had 32,297 total cases within the experimental period. The bottom barplot in Figure 5 denotes the area under the curve shown in the top line graph, totaling the number of cases per scenario.



[Fig. 5] US COVID-19 daily cases (top)
US COVID-19 total cases per scenario (bottom)

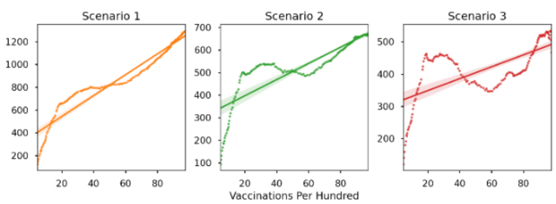
We calculated the efficacy of overall vaccination to be equivalent to the percent reduction from each scenario to the actual trend, thus meaning that the higher the percent reduction, the more effective vaccines were in that scenario. We calculated the percent reduction by comparing the values of each scenario denoted in the barplot to the actual trend. As seen in Table 1, Scenario 1 had the highest percent reduction at 79.12% and

Scenario 3 had the lowest at 63.66%.

[Table 1] Summary of case reduction due to vaccination

Assumed Scenario	Percent Reduction Due to Vaccination
Scenario 1	79.12%
Scenario 2	69.28%
Scenario 3	63.66%

We calculated the efficacy of overall vaccination to be equivalent to the percent reduction from each scenario to the actual trend, thus meaning that the higher the percent reduction, the more effective vaccines were in that scenario. We calculated the percent reduction by comparing the values of each scenario denoted in the barplot to the actual trend. As seen in Table 1, Scenario 1 had the highest percent reduction at 79.12% and Scenario 3 had the lowest at 63.66%.



[Fig. 6] Hypothesis Testing of US scenarios

Based on prior research conducted into the correlation of vaccinations and the number of daily cases, we hypothesized that in the three different scenarios studied in this paper, the number of individuals vaccinated would always have a significant impact in lowering the number of cases. We implemented hypothesis testing within the experiment to ensure that the quantified data was verified in multiple stages. The null hypothesis, H_0, was set to be that the vaccination does not reduce daily COVID-19 cases while the hypothesis, H_1, was set to be that the increase in vaccination rates would reduce daily COVID-19 cases.

[Table 2] R2 scores and p-values of experimental scenarios in US

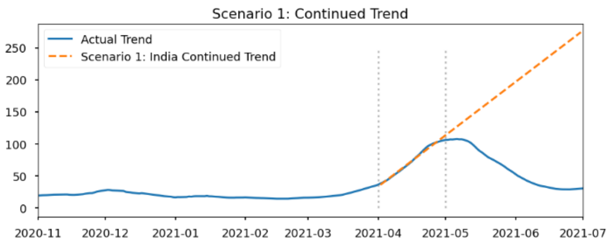
Assumed Scenario	p-value
Scenario 1	1.25e-77
Scenario 2	1.47e-34
Scenario 3	1.70e-4

The p-values played an important role in identifying efficacy. The formal definition of p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct. The low p-values, usually thresholded by whether it is greater than 0.05, indicate that the null hypothesis should be rejected, and that the outlined hypothesis should be accepted. Since all p-values are much lower than 0.05, we can claim that our hypothesis that vaccination rate indeed reduces daily COVID-19 cases, and by a statistically significant amount. Indeed, if the null hypothesis was true, then the slope of the projected line would have been closer to zero in Figure 6.

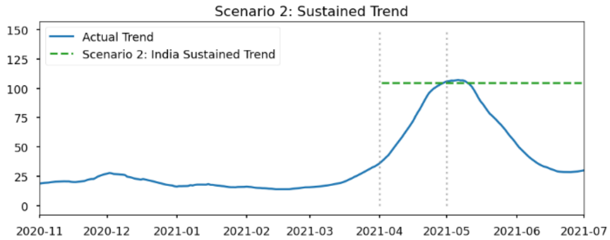
In the graphs found within Figure 6, the x-axis represents the percentage of people vaccinated within the US and the y-axis represents the reduced number of COVID-19 cases per million. As shown through the correlation of the x- and y-axes, there is a significant relationship between the two variables, indicating that vaccinations reduced the number of COVID-19 cases regardless of the scenario.

3.1 Additional Work

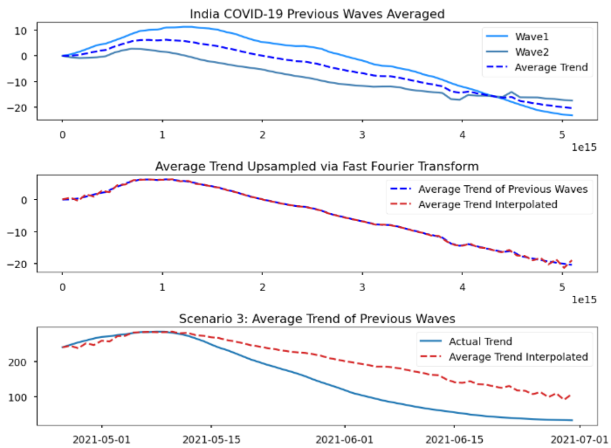
We applied the same three scenarios to the country of India in order to see the applicability of our data to other countries around the world. As seen in the above graphs and tables, the similarity in data is comparable to the results found from the data extrapolated from US cases of the coronavirus. The R2 scores from Table 3 show that all three scenarios are close to one, whereas the US study showed just two of the three scenarios to result in numbers close to one. This proves that our hypothesis is correct in setting a correlation between vaccination and decreased COVID-19 cases.



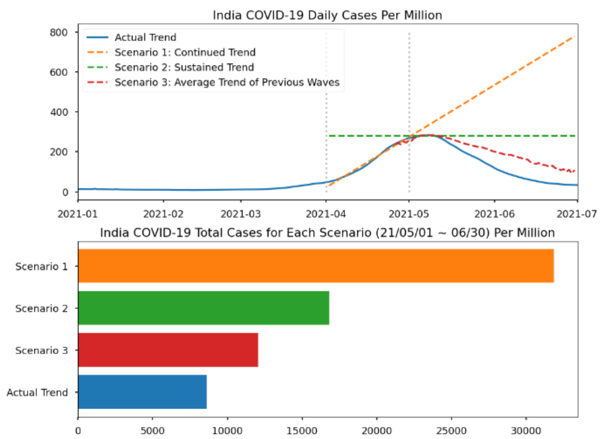
[Fig. 7] Trajectory of Continued Trend in India



[Fig. 8] Trajectory of Sustained Trend in India



[Fig. 9] Trajectory of Average Trend of Previous Waves in India

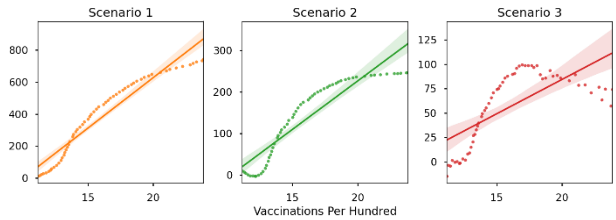


[Fig.10] India COVID-19 daily cases (top)

India COVID-19 total cases per scenario (bottom)

[Table 3] R2 scores and p-values of experimental scenarios in India

Assumed Scenario	p-value
Scenario 1	5.62e-65
Scenario 2	4.58e-41
Scenario 3	7.55e-17



[Fig. 11] Hypothesis Testing of India scenarios

4. Conclusion

The road to global vaccination has not come without difficulties. The first problem was the lack of supply during the height of the pandemic. In the United States, the vaccines were released in phases to healthcare workers and long-term care facility residents, essential workers, and finally to the general public aged 16 and older, even with this rollout plan in place, the production, delivery, and proper distribution of the vaccines could not keep up with the high demand by each phase.

The second hurdle came in the form of individuals who refused to be vaccinated. In first world countries, such as the U.S. or U.K., the pushback within society turned into a pressing social issue. Individuals voiced concerns about the potential side effects of the vaccines, lack of trust in the vaccines or their developers, or the fact that they did not believe COVID-19 to be a real threat.

In response to the lowered number of vaccinated individuals, corporations in the U.S. began offering incentives to those who were vaccinated, and also were able to present proof of their vaccination status. For example, Krispy Kreme offered free donuts, Staples offered free vaccination card lamination, and Lyft and Uber offered discounted rides to vaccination sites.

Furthermore, many companies have made it a requirement for employees to be vaccinated in order to continue employment. In hospitals throughout the United States, healthcare professionals who refused to be vaccinated were terminated from their positions. People who were previously regarded as heroes during the pandemic were left unemployed due to the same virus.

With the emergence of more threatening variants of the coronavirus, individuals who are eligible and able to be vaccinated should do so in a timely manner. The correlation between increased immunity and vaccination have been proven through the real-life cases studied in this paper and other reputable scientific source.

Though studies into the long-term effects of the vaccines and the ongoing trend of COVID-19 need to be continued, it is clear that building immunity towards a highly transmissible virus is the only way to return life to the way it was before the pandemic.

References

[1] Badiani, A.A. Patel, Jay. Ziolkowski, K. & Nielsen, Freja. Pfizer: The miracle vaccine for COVID-19?. Public Health in Practice. 2020,

[2] Terry, Mark. "Updated Comparing Covid-19 Vaccines: Timelines, Types and Prices." *BioSpace*, BioSpace, 14 Dec. 2021, https://www.biospace.com/article/comparing-covid-19-vaccines-pfizer-biontech-moderna-astrazeneca-oxford-j-and-j-russia-s-sputnik-v/.

[3] "How Covid-19 Is Changing the Approach for the Vaccine Development Process." *PPD Inc*, 20 Aug. 2021, https://www.ppd.com/blog/how-covid-19-is-changing-the-approach-for-the-vaccine-development-process/.

[4] Chodick, Gabriel & Tene, Lilac & Patalone, Tal & Gazit, Sivan & Ben Tov, Amir & Cohen, Dani & Muhsen, Khitam. The Effectiveness of the First Dose of BNT162b2 Vaccine in Reducing SARS-CoV-2 Infection: Real-World Evidence. SSRN Electronic Journal. 2021.

Appendix: Coefficient of Determination

If $\underline{y}$ is the mean of the data given by the following equation:

$$\underline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Then the variability of the dataset can be measured with two sum of squares formulas:

1. Residual sum of squares

$$SS_{res} = \sum_{i=1}^{n} (y_i - f_i)^2$$

2. Total sum of squares

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \underline{y})^2$$

The most general definition of the coefficient of determination is:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tol}}$$

In the case where a linear regression result has 0 residuals or in other words is a perfect fit, the residual sum of squares becomes 0, and thus the coefficient of determination becomes 1. In the worst-case scenario where the fit performs just as good as the residual sum of squares, the coefficient of determination becomes 0.

이재욱 (Jaweook Lee)

- Cheongshim International Academy
- 2020. 09. 14~ 현재

<관심 분야>

의료공학

함승준 (Alexander John Ham)

- Worcester Academy
- 2019. 08~ 현재

<관심 분야>

데이터 통계