Modified SIR Model for COVID-19 Pandemic in New York City

Xinyi Angela Cheng, Amelia Johnson Morrissey

May 2020

1 Abstract

As SARS-CoV-2 (COVID-19) continue to claim lives around the world, governments implement a variety of measures such as closing public beaches and ordering stay-at-home policies and conduct testings to minimize possible cases and deaths. At the same time, researchers strive to testing treatments and vaccines with the aim of fully controlling the spread of COVID-19. With curiosity of the future trend of the pandemic, we mainly focus on the epicenter in the United States – New York city and transform the SIR model with the goal of addressing data collection error. Our SUIRD model that we propose captures both the undetected number of infectious individuals and the number of deaths. We further identify three different cases for the undetected cases and then simulate the epidemic.

2 Introduction

Since the first case of COVID-19 was confirmed in Wuhan, China toward the end of 2019, its common-used name – "coronavirus" started to appear on the headlines of news and become a hot topics on social media [1]. Initially from China, the pandemic has quickly spread around the world through other Asia and Europe with Italy hit the earliest and hardest among all on the same continent. According to CDC (Centers for Disease Control and Prevention), the number of confirmed cases climbed up dramatically starting from mid-March when schools closed campus and switched to online coursework, and states government issued stay-at-home policy [2].

As of 6:00am CEST, 4 May 2020, there have been 3,349,786 reported cases and 238,628 deaths worldwide, of which 1,093,880 reported cases (32.66

%) and 62,406 deaths (25.15 %) have occurred in the U.S. [3]. The virus is spreading at an incredible rate; in the last 24 hours alone, there have been 26,753 new cases and 5,000 deaths in the U.S. alone, putting strain on healthcare systems. It is imperative to have accurate models in order to anticipate load healthcare systems.

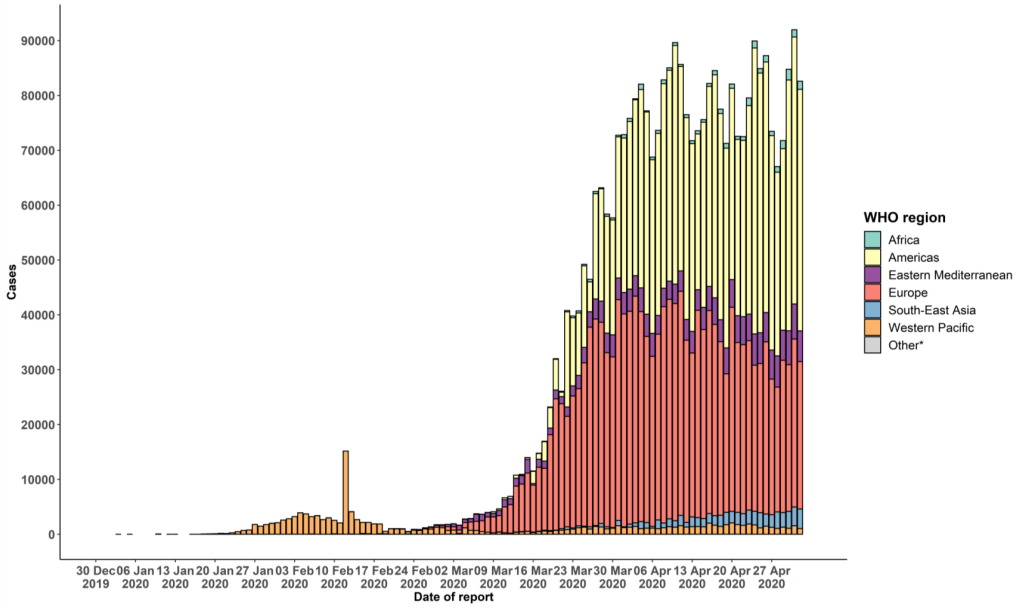

Figure 1: *Number of confirmed COVID-19 cases for each geographical regions defined by WHO from December 30, 2019 to May 3, 2020. As shown in the graph, there is a dramatic surge in Europe in early March and then in North America.[3].*

3 Background

The most commonly used models in the study of infectious diseases are SIR models as introduced by Kermack and McKendrick in 1927 [4]. From recent research papers regarding COVID-19 models, we find many applications of the base SIR model to the novel coronavirus that address some of these specifics of the current outbreak. Lin et al. and Yang et al. transformed the SIR model to a new SEIR model, adding an Exposed group in the population [1][5]. In "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions", the authors define $E(t)$ as a subpopulation that has been exposed to the virus but is not yet infected [5]. This population is still infectious but asymptomatic, and

therefore not considered 'infected.' This reflects the latency period wherein the virus is incubated but does not yet display symptoms.

This strategy is useful in accounting for population movement. However, it assumes that all of the asymptomatic exposed individuals move to $I(t)$ at the rate of incubation, i.e. all asymptomatic exposures become symptomatic. Similarly, this model neglects to address the testing insufficiencies in the U.S. by assuming that all infections are accurately reported as Yang et al. fit the data of reported cases together with the $I(t)$ curve to estimate R_0.

In reality, studies have shown that a significant amount of cases are never symptomatic or only present mild symptoms. According to New England Journal of Medicine (NEJM), over 50% of the residents from a Seattle-area nursing home has no symptoms at all when they were tested positive for COVID-19 [7]. Additionally, in the U.S., "limited test availability has led to largely restricting testing to those with more severe disease or those who are at risk of serious complications" [6]. This data collection error is especially prevalent in the U.S. as compared to other countries such as South Korea, which has "undertaken aggressive population-based screening and testing" [6].

Importantly, these 'hidden' cases, while unreported, are still infectious, and contribute to transmission of the disease. Unreported cases also make it difficult to implement public health measures that involve quarantining infected individuals. Moreover, SIR models and other modified SIR models that employ a simple linear structure (such as the SEIR model heretofore mentioned) fail to accommodate for a second groups of infected individuals who are undetected, decreasing the reliability of these models for predicting caseloads on public health systems.

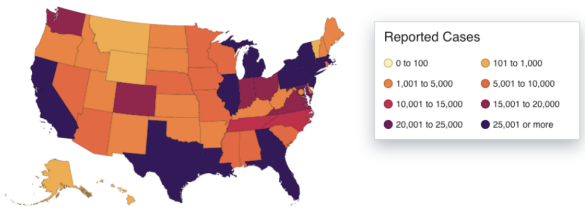

Figure 2: *A map of the U.S. from the CDC showing reported COVID-19 in each state.*

4 Methodology

4.1 Data

In this paper, we will look at the data from New York City, the current epicenter of the outbreak in the U.S with 166,883 reported cases and 13,156 deaths, 15.25% and 21.09% in the U.S., respectively. We source this data from the New York City Department of Health and summarize it below:

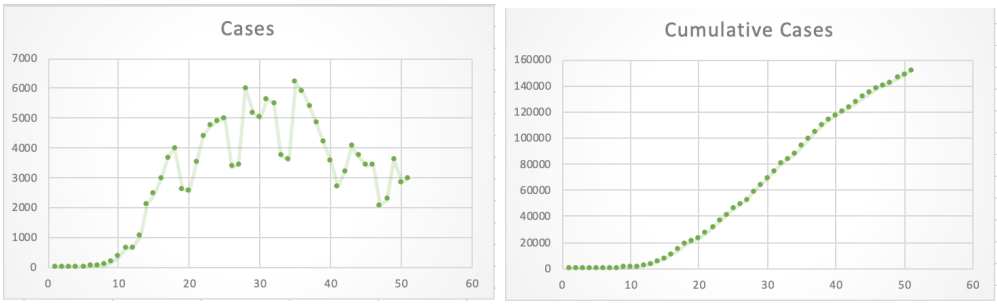

Figure 3: *Number of positive cases based on diagnosis date (left) and number of accumulative cases in New York city on a daily basis since March 3, 2020; data fit in Excel [8].*

When we apply the base SIR model, we see from figures below that our predictions do not fit the shape of the data well. We cannot achieve both the steep slope during the increasing period of the infection and the correct height of the infection curve.

For example, when predicting R_0 by assuming an exponential shape in the beginning of the infection and fitting the exponential model to the data we find an R_0 of 1.53 [9]. Applying this to a base SIR model we predict an infection peak much higher than then current peak infection load and cannot account for the fact that new case numbers are beginning to decline. Below, we compare the New York City data with the base SIR model with $R_0 = 1.53, 1.63$, and 1.73. In all three cases, we can visually see that the model does not fit the shape of the data well.

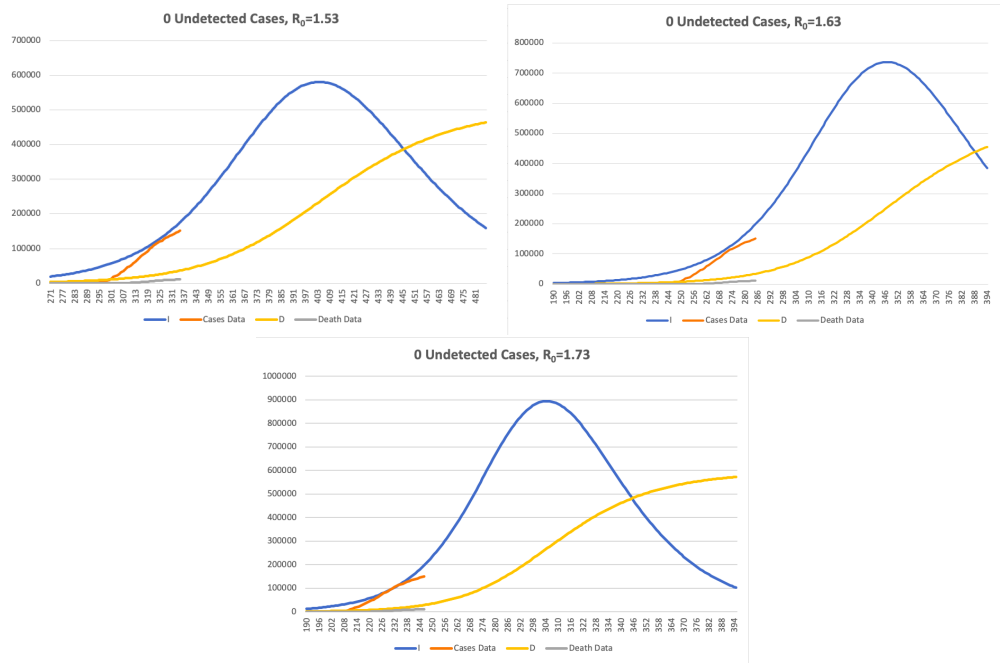

Figure 4: *We apply a base SIR model with various R_0 values, including the exponential model prediction of $R_0 = 1.53$, and compare them with reported cases and death data from New York City.*

Admittedly, a significant part of this reduction in the infection rate is likely due to public health measures. (Accordingly, we recommend the strategy proposed here be applied to a model with a time dependant Beta. See discussion.) However, data reporting error is still likely a major factor in this discrepancy. A recent study in New York City showed that 19.9% of New Yorkers possessed antibodies to the novel coronavirus, which signifies that they were exposed to the virus [10]. Here, we will propose a modification on the base SIR linear model that we believe could help to mitigate the error caused by testing insufficiencies and hidden asymptomatic infectious cases.

4.2 Model

4.2.1 Modification to SIR Model

As a common tool in disease modeling, SIR model partitions a large, closed population, N, into three compartments – the susceptible individuals, the infected individuals, and the recovered individuals. Each size of the subpopulation at time t is denoted as S(t), I(t), and R(t) respectively. In addition, it has two main assumption:

1. Mass action mixing, that is, the rate of interaction between the susceptible and the infectious is proportional to the product of S(t) and I(t).

2. Recovery rate of the infected is proportional to the size of the infectious population.

To understand the trend of epidemic, epidemiologists utilize the reproductive number, R_0, to calculate how many new infections are caused by an earlier infection. The estimate of R_0 is below 1.0 wherever the number of cases is going down and above 1.0 wherever the number of cases is going up.

Based on the SIR model, we modified it and built a new SUIRD model. First, we separated the infection population into the detected and undetected one by taking account into the fact that a significant number of cases may not be captured by the current data. In addition, we added one more subpopulation representing death, which has more accurate and reliable data than the case data [6].

4.2.2 SUIRD Model

We specifically focused on modeling the epicenter of COVID-19 – New York city, with a population size of 8.399 million people. Our model assumes a closed population, N, as SIR model and divides it into five compartments instead – the number of susceptible population, the undetected number of infectious, the detected number of infectious, the number of recovered, and the number of death, each denoted as follows such that $N = S(t) + U(t) + I(t) + R(t) + D(t)$:

- N = total population size of New York city.
- $S(t)$ = the size of susceptible individuals who can get sick at time t.
- $U(t)$ = the size of undetected infectious individuals who are sick and contagious to the susceptible at time t.
- $I(t)$ = the size of detected infectious individuals who are sick and contagious to the susceptible at time t.
- $R(t)$ = the size of recovered individuals who are immune and cannot get infected again at time t.
- $D(t)$ = the size of removed or dead individuals at time t.

We describe the rate of change of the size for these five groups using differential equations over time, t, with opposing signs based on whether the change is moving individuals in or out of a particular group.

$$\frac{dS}{dt} = -\beta S(I+U) - \beta r S(I+U) \tag{1}$$

$$\frac{dU}{dt} = \beta S(I+U) - vI \tag{2}$$

$$\frac{dI}{dt} = \beta r S(I+U) - vU \tag{3}$$

$$\frac{dR}{dt} = vI(1-f) + vU \tag{4}$$

$$\frac{dD}{dt} = vIf \tag{5}$$

where $\beta > 0$ is the transmission rate, $v > 0$ is the recovery rate, $r \geq 0$ is the ratio of the undetected cases to the detected ones, and f is the fatality rate of the detected infectious individuals.

Similar to the SIR model, the transmission rate (β) is based on the strength of the pathogen in transmitting, called transmissibility of the infectious disease, τ, and the amount of contacts, k, each infected individual has per unit time. Therefore, $\beta = \frac{k\tau}{N} = \frac{b}{N}$, where $b = k \cdot \tau$. We assume a recovery period of 14 days for infected people across all age groups based on reports by John Hopkins University [11].

In addition to the transmission rate and recovery rate, we added another two parameters, r and f that do not change over time. Assuming the total infectious population consists of two groups, we utilized to the result of antibody test conducted in New York city to estimate the ratio of the undetected infectious people to the detected ones such that $r = \frac{p \cdot N}{I_t}$, where p is the percentage of population with COVID-19 antibodies and I_t is the number of detected infectious people by the latest date in the original dataset. As mentioned previously, the antibody test suggests 19.9% of the population in New York have developed antibodies for COVID-19. We assume that all the individuals with antibodies are the undetected population of infectious.

Based on the total confirmed cases and deaths by the end of April, we set the fatality rate to 10%, excluding the probable death. According to New York city government, a probable death is caused by "COVID-19" or an equivalent but has no known positive laboratory test for SARS-CoV-2 (COVID-19) [12].

4.2.3 Estimate Transmission Rate

To estimate β, we first visually fit the model to the data by changing R_0 by degree until the shape was the best fit, but keeping the model peak fixed on our estimated peak time of 4/1/2020. Changing our estimate of R_0 will not change the real peak of the outbreak. It would change the estimate of the total length of the outbreak, which is reflected in longer tails before and after the peak. Then, we used the following equation derived from SIR model to calculate β:

$$\beta = \frac{k\tau}{N} = \frac{b}{N}$$

Additionally, we calculate an adjusted β, written β_a:

$$\beta_a = \frac{\beta}{r+1}$$

where $r = \frac{p \cdot N}{N}$ is the ratio of the number of individuals with antibodies to the total population.

β_a can be interpreted theoretically as the portion of the infection rate contributed to the cases that were reported. We weight the contribution of the reported and undetected infections to the transmission of the virus based on the relative sizes of each group. We give undetected infected subpopulation, U, a weight of r and detected infected subpopulation, I, a weight of 1. β_a, then is the base infection affect unit given to one unit of weight.

5 Results

We calculate all the metrics, again for New York City, with three estimates of the percent of the population possessing antibodies: 15%, 20% and 25%. We use a time step of 24 hours.

p	0.15	0.20	0.25
t	1.00	1.00	1.00
R_0	2.20	2.40	2.60
B	0.0000000187	0.0000000204	0.0000000221
B_a	0.0000000020	0.0000000017	0.0000000015
r	8.32	11.09	13.87
v	0.07	0.07	0.07
N	8399000.00	8399000.00	8399000.00
f	0.10	0.10	0.10

Figure 5: *Table of estimates for eight parameters defined in Model section with three different cases.*

We note that there is a linear relationship between the believed percent

of people with antibodies and our estimate of R_0. This shows that our visual fitting of the model to the data to estimate R_0 was consistent.

Using the estimated parameters shown in Figure 5, we fit the following curves (*Figures* 6, 7, and 8) in Excel, and find the epidemiological metrics below:

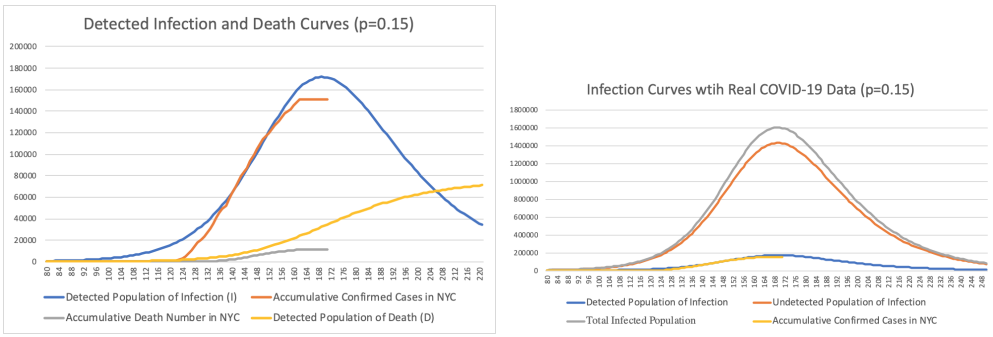

Figure 6: *Real data fitted for number of deaths and confirmed cases with 15% of the total NYC population as undetected.*

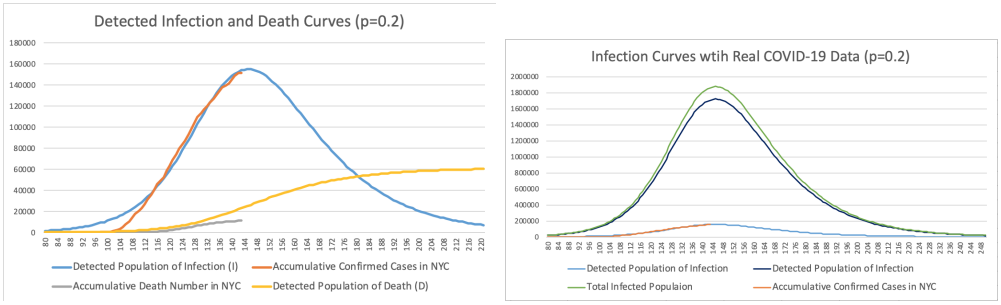

Figure 7: *Real data fitted for number of deaths and confirmed cases with 20% of the total NYC population as undetected.*

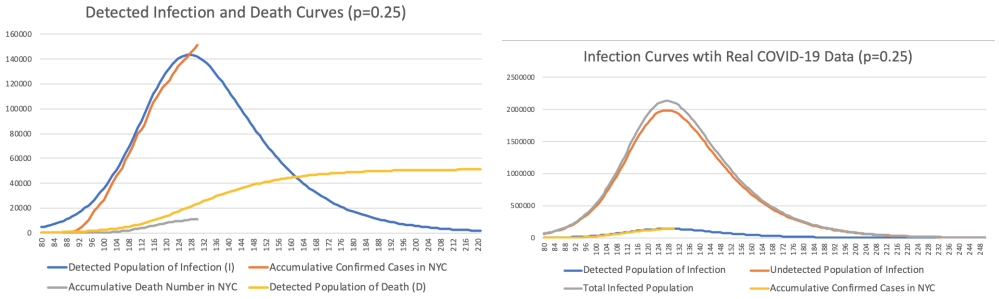

Figure 8: *Real data fitted for number of deaths and confirmed cases with 25% of the total NYC population as undetected.*

p	0.15	0.20	0.25
I max	155109.28	171863.00	143405.18
U max	1720534.35	1429782.10	1988384.85
I + U max	3054536.62	2978467.29	3069590.22
detected size	613285.84	763953.01	513789.39
percent detected size	0.07	0.09	0.06
undetected size	7417212.68	7122134.30	7638547.33
percent undetected size	0.88	0.85	0.91
Date of peak	4/1/20	4/1/20	4/1/20
death toll	61328.58	76395.30	51378.94

Figure 9: *Table summarizing results for three different cases for the number of undetected infectious people with antibodies.*

The most important results that public health officials attempt to estimate in order to predict load on healthcare systems and on the public are the maximum infections at one given time (I_{max}), the size of the epidemic (the amount of infections over the entire course of the epidemic, and the death toll. We see from our results that the maximum estimate for all three of these metrics occur when greater than 15% but less than 25% of the population has antibodies. When the undetected curve is smaller (15%), This means that there are less infections at any given time to transmit the virus to more people.

On the other hand, when the undetected curve is larger (25%), the virus spreads more quickly, causing a sharp decrease in the susceptible population (as shown in Figure 11), which in turn lessens the blow of the infection, assuming that possessing antibodies correlates strongly with at least short term immunity.

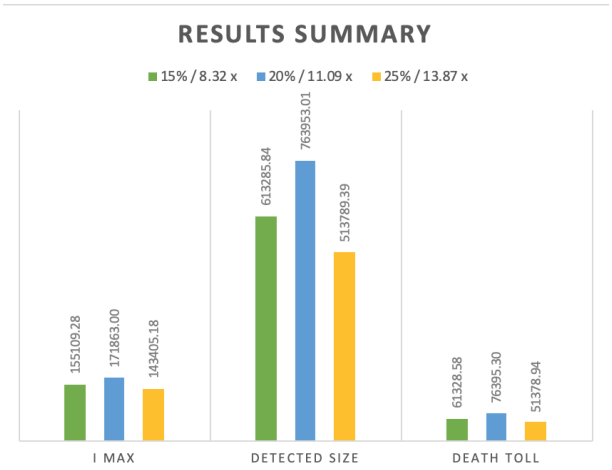

Figure 10: *Graph summarizing the maximum number of detected infectious people (I_{max}), the detected size of the epidemic ($I(\infty)$*, and the total number of deaths in three cases; the label for each color represents the percentage of people with antibodies/ratios of undetected infectious individuals and the detected.

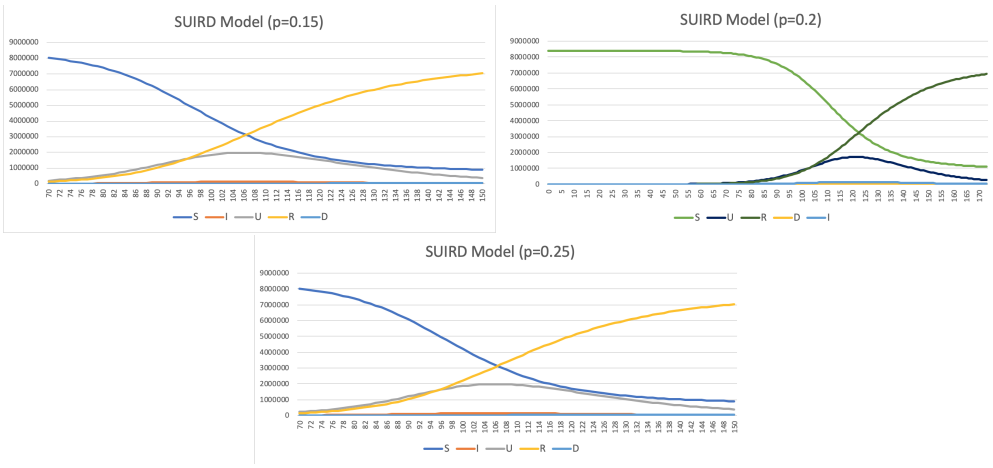

Figure 11: *SUIRD model simulations for three different cases.*

6 Discussion

We build a modified SIR model targeting the quality of data collection. Based on the result of the most recent antibody study released on May 2, 2020, we analyze three different cases in terms of the number of undetected infectious individuals. By including the undetected group of infectious individuals and the number of deaths, our SUIRD model mitigates the error in

confirmed cases in New York city.

Previous works related to COVID-19 have made projections in terms of the number of cases and deaths on a daily basis as well as the excess amount of medical supplies needed such as ICU beds and ventilators [6]. Piccolomini et al. propose a Susceptible-Infected-Exposed-Recovered-Dead (SEIRD) Model and take account into the impact of Italian government politics on the pandemic with a time-dependent transmission rate [13]. However, none of them suggest any solutions to the inaccurate data as SUIRD model.

As shown in the previous section, the estimation of R_0 seems to fit the SUIDR model very well, but other parameters used in the model largely depend on assumptions that may not apply to the reality. As we assume all people with antibodies were infected at some point and recovered, the ratio (r) of the undetected infectious people to the detected ones may not be accurate since there is little evidence that suggests people with COVID-19 antibodies to have immunity to the virus forever [10]. Even for the people who have recovered from COVID-19, they are not guaranteed immune. In this case, our estimation serves as a baseline and requires further study for improvement in SUIRD model.

In addition, as we assume the average recovery period to be 14 days, the SUIRD model does not consider different recovery rates for different age groups. According to the study reported by John Hopkins University, the age group does significantly affect how long it takes for one to recover from the coronavirus.

The SUIRD model provides a good fit of the real data from New York city's COVID-19 confirmed cases and deaths. It establishes a fundamental framework for other disease models to help mitigate the error caused by data collection and achieve a more accurate prediction. Possible future work include a time-independent infection rate as suggested by Piccolomini et al. and a further analysis on the impact of the time when to lift up government measures on the public health using SUIRD model.

References

[1] Lin et al. *A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action* (2020). International Journal of Infectious Diseases 93 (2020) 211-216. https://doi.org/10.1016/j.ijid.2020.02.058

[2] Centers for Disease Control and Prevention. (2020). *Cases in the U.S.* https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html?fbclid=IwAR2YGdSiJ1zk6mktakCLsCqjU-tEq9XsvLMK2fGG0vmHPIsAdMgl8C13cOU

[3] World Health Organization. (2020). *Coronavirus disease (COVID-19) Situation Report – 104.* 15 pages. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200503-covid-19-sitrep-104.pdf?sfvrsn=53328f46$_2$

[4] Weiss, Howard. *The SIR Model and the Foundations of Public Health* (2013). MATerials MATemàtics Volum 2013, treball no. 3, 17 pp. ISSN: 1887-1097 Publicació electrònica de divulgació del Departament de Matemàtiques de la Universitat Autònoma de Barcelona www.mat.uab.cat/matmat

[5] Yang et al. *Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions.* Journal of Thoracic Disease, Vol 12, No 3 (March 2020), 165-174. http://dx.doi.org/10.21037/jtd.2020.02.64

[6] Murray, C. (2020). *Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator- days and deaths by US state in the next 4 months.* The Institute for Health Metrics and Evaluation (IHME). https://doi.org/10.1101/2020.04.21.20074732

[7] Mary Van Beusekom. *Study: Many asymptomatic COVID-19 cases undetected.* Retrieved May 3, 2020 from https://www.cidrap.umn.edu/news-perspective/2020/04/study-many-asymptomatic-covid-19-cases-undetected

[8] NYC Health. *COVID-19: Data.* Retrieved April 22th, 2020 from https://www1.nyc.gov/site/doh/covid/covid-19-data.page

[9] Aaron A. King. (2017) *Introduction to inference: parameter estimation* Retrieved April 30, 2020 from https://kingaa.github.io/clim-dis/parest/parest.htmlfeature-based-parameter-estimation.

[10] Orion Rummler. *New York releases preliminary coronavirus antibody test results.* Retrieved May 2, 2020 from https://www.axios.com/coronavirus-new-york-antibody-test-f4fbed78-646f-4b46-90b8-5e8ca75380e4.html.

[11] Jonathan Eichberger. *Study of Data from Shenzhen, China, Provides Key COVID-19 Insights Analysis.* Retrieved April 28th, 2020 from https://hub.jhu.edu/2020/04/28/shenzhen-cdc-coronavirus-study/

[12] NYC Health. *Confirmed and Probable COVID-19 Deaths Daily Report.* Retrieved May 4, 2020 from https://www1.nyc.gov/assets/doh/downloads/pdf/imm/covid-19-deaths-confirmed-probable-daily-04142020.pdf

[13] Piccolomini, E., and Zama, F. (2020).*Monitoring Italian COVID-19 spread by an adaptive SEIRD model.* DOI: https://doi.org/10.1101/2020.04.03.20049734