5-4-2022

Epidemic Vulnerability Index: Vaccine Dissemination Criteria for Successful Resolution to Epidemics

Hunmin Lee

Epidemic Vulnerability Index: Vaccine Dissemination Criteria for Successful Resolution to Epidemics

by

Hunmin Lee

Under the Direction of Yingshu Li, PhD

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2022

ABSTRACT

Vaccination is the preventative measure that effectively decelerates the virus proliferation in a community. A successful response strategy toward pandemics can be obtained through selecting the optimal vaccine distribution route and minimizing the casualties by lowering the death rate and infection rate. In this thesis paper, we propose the Epidemic Vulnerability Index (EVI) that quantifies the potential risk of the subject via analyzing the COVID-19 patient dataset that correlates with mortality and social network analysis that affects the infection rate. We propagate the virus and vaccination in an Agent-based model based on real-world statistics of physical connections and features to 300,000 agents with nine vaccination criteria, including EVI. Vaccination through descending order of EVI has shown the best performance with the numerical outcome of 5.0% lower infection cases, 9.4% lower death cases, and 3.5% lower death rates than the average of other vaccination dissemination criteria.

INDEX WORDS: COVID-19, Data analysis, Dynamics of Vaccine & Virus Proliferation, Epidemic Vulnerability Index, Simulation-based inference, Stochastic proliferation estimation, Vaccine distribution strategy

Epidemic Vulnerability Index: Vaccine Dissemination Criteria for Successful Resolution to Epidemics

by

Hunmin Lee

Committee Chair: Yingshu Li

Committee: Zhipeng Cai

Yanqing Zhang

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

May 2022

ACKNOWLEDGEMENTS

First of all, I would like to thank my academic advisors: Dr. Yingshu Li and Dr. Zhipeng Cai, for their delicate guidance and support. Although I was initially from another advisor, they took me in as their student and gave me diverse opportunities without any conditions. I learned so much from them, in both research and how to build successful relationships. They are my role models, and I personally wish to be like them in the future. I would also like to thank my fellow students: Yueyang Liu and Avais Jan, who I co-worked with, and who provided me smile during my research. It has been almost two years since I started my study in the Department of Computer Science, and every faculty, staff were very understanding and considerate. Therefore, I am sending them my thanks. Finally, I would like to thank my family, especially my parents, for their continuous support.

Hunmin Lee

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

In late 2020, a novel virus COVID-19 (SARS-CoV-2) [1,2] has been proliferated and deranged the world, changing the diverse scene of the modern systems. Hundreds of variant types have been reported to academia and have been dispersed by mutation inside the host (Alpha (2020.09.03), Beta (2020.09.01), Delta (2021.03.01), Delta+ (2021.03.31), Omicron (2021.11.12), Stealth Omicron (2021.11.17), Deltacron (2022.01.09), etc.), rapidly escalating the infection and death cases in worldwide level. Scientists and researchers have successfully invented the COVID-19 vaccines, and vaccines were inoculated at an urgent pace. However, due to the increasing demand and the limited supply, insufficient resources were given to the governments. Furthermore, more and more variants have appeared, simultaneously increasing the inoculation doses of vaccines, and many nations suffer from severe deficiency of vaccines. This is a common phenomenon, especially during the preliminary stages after the outbreak, and while new variants are constantly being propagated, the optimal distribution of vaccines should be highly encouraged. Due to this situation, an effective vaccine distribution strategy must be established through multi-perspective analyses of virus dissemination dynamics and COVID-19 statistics. The main objective is to successfully gain control via decreasing the current infection rate and death rate by determining proper vaccination routes followed by other preventative measures.

This thesis paper presents the Epidemic Vulnerability Index (EVI), which quantitatively calculates the potential risk of an individual subject through its unique clinical and social factors. Internal clinical factors such as age, gender, underlying diseases, etc., have statistically shown the relationships with the level of factors, indicating the potential mortality after the subject is infected. Various research was reported concerning the correlation between the internal characteristics of the COVID-19 patient and its impact, measuring the severity of the disease for the subject's health

levels. Clinical factors indicate the risk level, which is quantifiable by a statistical dataset from past COVID-19 patients. The social factor is an external property that directly affects the infection rate compared to the clinical factor. By assessing the importance of the subject among its affiliated community, it offers the information of quantifying the potential risk of being infected. By implementing social network analysis, we compute the centrality of the nodes that compose the agent-based graph network. Discovering the nodes considered to have a relatively significant impact that accelerates the infection in advance is critical, as evading those nodes will appreciably lower the casualties. By collecting and analyzing the statistical attributes of the dataset involving past COVID-19 patient and their social network, this thesis numerically organizes the potential effect of internal factors and external factors through multiple simulations.

Through the EVI's risk assessment, we experimentally search the effective vaccination route based on the graph network by leveraging the EVI. Bringing the pandemic to an end means minimizing the death cases that were already infected, and gradually the number of infected hosts converges to zero. EVI is specifically designed to estimate the degree of possible mortality and infection rate by incorporating the unique factors of each subject; thus, considering EVI when devising vaccine distribution strategy is suitable. In order to show the viability, we design the simulation with an Agent-based Model (ABM) for validation. Our constructed ABM is graph-structured with 300,000 agents (subjects; nodes) and its physical interconnection (directed edges). In an ABM, the nodes indicate the hosts that the virus inside the community can infect. The nodes are statistically allocated with the designated clinical factors and set the edges based on their age. Edges represent the physical contacts, and the number of edges is determined based on the number of physical connections under the geographical constraints. During the initial steps, all nodes start with non-infected. Our simulation randomly selects the number of initial spreaders and injects the

vaccines into the specific targets through the nine simulation criteria. We adjust the parameters to observe the variation and outcome of infection cases, death cases, and death rates as the virus spreads throughout the community.

Contributions of this thesis can be narrowed down to three categories:

1. This thesis proposes the Epidemic Vulnerability Index that numerically estimates the potential risk of the subject, primarily focusing on quantifying the risk of death rate (clinical factors) and the infection rate (social factors) through statistical analysis associated with corresponding attributes. By fine-tuning the parameters, EVI could be optimized in given scenarios and heterogeneous settings.

2. This thesis suggests the stochastic proliferation simulation algorithm in the predefined ABM. ABM is constructed under real-world statistics, which allocates the embedded features with statistical properties of real-life to enhance confidence. Simulation-based ABM offers estimation and inference towards making prognosis of future pandemic processes. Also, it serves as a reference model for establishing an effective vaccine distribution route through predictive analysis.

3. By conducting nine vaccine distribution simulation scenarios, including the previously suggested metrics such as CVI [3], SVI [4], and PVI [5], this thesis offers the comparison of three metrics among the ABM: infection cases, death cases, and death rate. The results show that vaccination through descending order of EVI is shown to have the lowest value in those metrics.

4. We suggest the practical analysis based on simulation results, interpreting the difference between the COVID-19 statistics shown until the present time, and explicating the valid vaccine dissemination approach to inaugurate an effective impact.

Predicting and modeling the dynamics of pandemic proliferation is a complex mathematical task, for it is not clearly defined whether it is a deterministic process or a

probabilistic process. Furthermore, the estimation accuracy may not be robust for intrinsic features of the existing community comprehensively varies, and fine-tuning the parameters for regional prognosis approximation is required. In addition, parameters should not be static with respect to time (*i.e.*, random walk) but should dynamically be changed. Moreover, new variables are constantly emerging with the occasional period not only in statistics concerning clinical & social attributes but in cross domains such as political agenda, economic affairs, natural incidents, locational environments, etc. Therefore**,** an accurate prediction along with estimating a vaccine distribution impact with diverse circumstances is an uneasy task. In such a convoluted setting, possible results based on a simulation model through multiple empirical results offer us greedy-based options. It directs the value of coordinating the parameters in the search space, although a global minimum is not guaranteed. In addition, conducting data analysis aids when discovering the shape of the search space, providing insights and references. Likewise, this thesis presents the possibility of options when deriving the optimal solutions, especially with respect to locating the optimal path for vaccine dispersion to minimize death cases and infection cases compared to the existing vaccine distribution method and other pre-defined indexes. Through a comprehensive analysis of the past COVID-19 patient dataset among the 50 States of the United States, we aggregate the independent risks of each feature with respect to infection rate and death rate. We normalize the values through the linear combination of the risks based on multiple feature criteria. With the ABM consisting of 300,000 agents with the real-world statistical dataset, we experimentally spread the virus in our diffusion algorithm as well as vaccines with nine criteria, including EVI, and observed the outcome of infection cases, death cases, and cured cases, and agents with no virus. Finally, we evaluate the numerical ramifications of each vaccination scenario and validate our assertion.

This thesis paper is organized as follows. In section 2, the precedented researches were explicated in three sub-sections: Vaccine distribution in epidemiological cases, risk evaluation in pandemics, and the SIR model. In section 3, we present the mortality computation based on the clinical statistical dataset of COVID-19 patients in the US. Section 4 calculates the infection rate with the social statistical dataset and designs the simulation and the ABM analogous to reality for accurate dynamics. Section 5 introduces the proliferation algorithm and nine vaccine distribution simulation results, comparing the infection, death cases, and death rates that validate the impact. In addition, we suggest interpretations of the empirical result, followed by reflecting the explanation of the real-world COVID-19 statistics. Finally, we conclude this thesis paper by proposing feasible future applications and limitations in the real-practice concerning the current pandemic trend.

2 RELATED WORKS

According to the World Health Organization (WHO), the novel virus that incurs fatal diseases has constantly emerged, threatening the world's safety. Several global-scale pandemics have struck the world throughout the past two decades, such as SARS, Ebola, and COVID-19. Accordingly, epidemiological studies that evaluate the potential risk based on the statistical biological dataset were conducted, suggesting the index that assesses the relative risk [3-11]. These indices serve as a reference model that can be considered when making virus-preventative strategies [12-16], which helps as an auxiliary method for modeling more comprehensive and accurate virus dynamics. Furthermore, strategic approaches to allocating medical supplies such as vaccines, masks, sanitizers, and medicine in a suitable time and place have been studied [17-20] concerning the current environmental status (*e.g.*, regional factors, infra status, etc.) for an effective propagation. The previous research that defines the index that estimates the risk primarily concentrates on the clinical risk, which considers the internal biometric attributes of the patient infected by the target virus. However, this thesis paper presents the risk that aggregates the infection rate and vaccine distribution effect based on death rate and infection rate. This section is composed of three subsections. In the first subsection, related works concerning vaccine distribution methodology were introduced. In the second subsection, past studies evaluating the internal risk when the host is infected were explained. Finally, the SIR model analogous to the vaccination distribution simulation is analyzed in detail in the final subsection.

2.1 Optimal Vaccine Distribution

Among the various preventative measures against pandemics, the vaccine is a medically validated option and thus recommended to contain the virus dispersion among the community

effectively. The most effective way is to inject the vaccines [21] into every possible host, but the limited production is a constraint that hinders while the infection is an ongoing process. This phenomenon is evident, especially during the initial phase after the vaccine starts being propagated. Thus, an effective diffusion scheme is crucial through diverse analysis to make the best use of the existing resources.

Many studies assume that the epidemic proliferation process is deterministic and develop models using Ordinary Differential Equation (ODE). On the other hand, research is underway to construct probabilistic predictive models. Matrajt et al. [22] suggested evidence-based guidance that shows the effect of the vaccination route through empirical studies. It focused on the analysis of vaccination by age group throughout the time period. Various empirical studies that illustrate the vaccination effect were suggested in order to discover the optimal vaccination trajectory. Frequently, the vaccination process conducted the simulations on the Agent-Based Model process [23, 24], which is composed of unique entities (agents), and their stochastic interaction triggers the dissemination of the virus throughout the community model. Kerr et al. [15] proposed the agent-based COVID-19 simulator that can be adjusted to specific locational properties and hyperparameters, which offers adaptability to a practical scenario. Silva et al. [25] devised seven social activities of the designated regional community. This research correlated with the economic impact and the death cases in the predefined ABM. Studies that formulated AGM under regional constraints that incorporate the indigenous characteristics were suggested, such as the city of Bogatá, Columbia (Gomez et al. [26]), and New York City, USA (Hoertel et al. [27]). Subsequent to constructing a real-world-based ABM structure, the predictions and measuring of the variations of casualties and other factors have been studied via simulations using disease dynamics. Hinch et al. [28] designed the traceable ABM model based on the COVID-19 patients of England with the

age-stratified patient's physical contact dataset. In a similar context, one of the criteria that this thesis implemented the analysis was the age-stratified groups since the open dataset that provides the information of physical contact per time unit that we collected was evidently categorized by age groups. Furthermore, interconnection via age group is a decisive criterion that has been validated [29] as a useful measure. The recent vaccination distribution studies' major limitation is that it only concentrates on the criteria of age-stratified group feature. This is due to most of the COVID-19 open-source datasets being assorted via age groups; however, this thesis endeavors to collect more datasets concerning other features. Apart from the age-level feature, we consider 22 types of underlying disease, centrality, locational factors, and gender to enhance the accuracy of scoring the potential risk or EVI. Through this estimation, we implement the EVI for designing an effective vaccine distribution route that triggers a lower number of cases of infection and death in the ABM assembled from real-world statistics.

2.2 Risk Evaluation based on Statistical Data

As previously described, various studies were conducted to accurately assess possible dangers that the subject may involve after the host was infected. This is considered critical research for categorizing the degree of risk will lead to effectively determining the target set, which will broadly impact the number of casualties. These studies measured the future statistics based on the current risk scenarios for successful countermeasures to stagnate the negative impact. These studies were conducted not only in disease proliferation but other natural disasters. For example, The Centers for Disease Control (CDC) proposed the universal index (Social Vulnerability Index, SVI) [4] to quantify the potential vulnerability when exposed to a natural disaster. It was initially designed based on the accounted damage of the casualties from Hurricane Katrina in 2005. This

served as an impactive criterion to classify the social vulnerability and the risk that may follow when exposed to danger, as the socially vulnerable are highly prone to be adversely affected. Likewise, weighing the vulnerabilities with the age relation has been indexed (Frailty Index, FI [9]), measuring the relative status of physical and mental health. Several studies were introduced that discovered the relationships between FI and the COVID-19 risk [10, 11], primarily focused on elderly citizens. However, some research asserts that the FI is not sufficiently suitable to define the internal threat of COVID-19 patients. Based on the COVID-19 patient diagnosis rate in the state of Washington in the US, Amram et al. [6] graphically visualized the geographical maps that offer intuitive statistics that are plainly opened to the public. It displays the county-level risk and the current statistics of COVID casualties of infection and death cases. Decaprio et al. [3] suggested the COVID-19 index (C-19 Index), which computed the risk, utilizing the XGBoost machine learning algorithm via training the respiratory disease patient dataset. The respiratory patient shares similar clinical symptoms to the COVID-19 patient and is linearly proportional to the death cases of the COVID-19 patient, but implementing another target dataset lacks credibility when adopting the response strategies. Marvel et al. [5] proposed a similar index: Pandemic Vulnerability Index (PVI), which calculates the risk scores with the Bayesian model by training the US county-level COVID-19 patient dataset. The authors also presented the dashboard of statistical visualization in a geographical display that shows the numerical casualties. Previously suggested indexes mainly were trained through the county-unit risk related to death, offering the current risk scores of the county [30]. Furthermore, studies were presented that analyzed the economic and financial impact and quantified the degree [8], such as Global Fear Index [7]. Most of the studies are concentrated on identifying the hidden threats when the host is infected; thus, only clinical factors were taken into account. Also, geographical units such as state-level and

county-level death rates were statistically verified, which is also related to clinical aspects. This thesis takes further steps, considering biological and social factors, covering the death rate and infection rate simultaneously. Since infection rate and the death rate have a negative linear correlation in the perspective of entity features (which will be further explained in section 4~5), both factors must coexist when distributing vaccines in order to eradicate the virus in the community successfully. By optimally adjusting the contribution weights of each attribute depending on the current situation, establishing the adaptive measure for effective response can be obtained.

2.3 SIR Model

SIR model is a well-known and widely used mathematical model that was proposed by William. O. Kermack and A. G. McKendrick (1927), and along with the basic reproduction number (R_0), it has been a fundamental model that has been applied to prognose the variation of a number of the Susceptible, Infectious, and Recovered throughout the time. R_0 offers the quantitative measurement of the number of infections from the patient zero. When $R_0 > 1$, it becomes a pandemic, spreading the infection exponentially during the initial steps. Additionally, the virus that shows $R_0 \approx 1$ is known to be endemic that spreads only in a specific region. On the contrary, the infection cases decrease when $R_0 < 1$, the virus perishes from the community, with every host recovered. The epidemiologists from the seven authoritative medical institutions have declared that the R_0 ranges from 2.2 to 3.3 (2020. May), which is sufficient to be a pandemic. The Spanish flu (H1N1 virus) influenza from 1918 had approximately 1.8 R_0 score, SARS (Severe Acute Respiratory Syndrome, 2002) had 3.0 R_0 score, and the MERS (Middle East Respiratory

Syndrome, 2012) had 0.8 ~ 1.3 R_0 score. MERS was known as an epidemic, which entails a lower R_0 value compared to the pandemic virus.

R_0 is defined through the aggregation of the following terms: $infection\ rate\ (\tau) \times contact\ interval\ (k) \times duration\ period\ (\delta)$. Mostly, τ and δ approximates to the constant (*i.e.*, $\max(\delta) \approx 14\ days$), whereas k may differ by surrounding factors such as regional population density, patient's trajectory, etc. When R_0 explains the numerical value for initial patient zero, another index called Effective Reproduction Number was suggested R_E which implies the average of infection cases from the patient. R_E differs from the R_0 from the fact that it may vary from the factors that hinders the virus transmission such as vaccination or other preventative endeavors. Moreover, R_T was proposed which indicates the R_E in certain time period T, and most studies refer to the combination of R_E and R_T as R [31], such that $R \neq R_0$. In order to decrease the virus proliferation and bring an end to the infection towards the susceptible is to discover an effective scheme to make $R < R_0$. To elaborate, infection cases from R should gradually be lower than initial R_0, until $\lim_{t \to T} R_t < 1$ where T is a time period. Thus, government and health authorities invest their full efforts with respect to decrease the value of τ, k, δ. A well-known method to decrease τ is to encourage washing hands and wear masks, and there are measures such as quarantines for the infected and social distancing to diminish k. Hospital treatment and enhancing immune system would minimize δ. Let $w_\tau \cdot \tau + w_k \cdot k + w_\delta \cdot \delta = c$, which connotates the ratio of prevention. The R can be defined with the following function $R = R_0(1-c)(1-p)$ where p is the population ratio (%) who maintains the immune system towards the virus. This informs that increasing the c, p will impede the spread. Assume that $c = 0$, and we have $R = R_0(1)(1-p)$. Let $R_0 = 3$ and our objective is to become R to be lower than 1. In this case, $1 < 3(1-p)$, and $p > \frac{1}{3}$, which implies that the $\frac{1}{3}$ of population is infected, it will downturn

the infection having $R < 1$. However, R model is inferior to the super spreader; the spreader who infects others with much larger than R or R_0, and this indicates that the R is susceptible to outliers. SIR model is composed of three following terms: S, I, R, each indicating number of susceptible, infected, and recovered, respectively. $S + I + R = N$, where N refers to the number of populations in the target community. Through derivatives of each term by time *t*, we know that $\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$, also $\left|\frac{dS}{dt}\right| = \left|\frac{dI}{dt}\right| + \left|\frac{dR}{dt}\right|$. Note that when $\frac{d^2I}{dt^2} > 0$, the number of patients accumulates exponentially, and $\frac{d^2I}{dt^2} < 0$ refers that the increasing speed of accumulation are decreasing. Furthermore, $\frac{dI}{dt} = \beta SI - \gamma I$, where β indicates the probability of contact resulting infection, and γ is the recovery time such that $\gamma = \frac{1}{period\ of\ infection}$. $\gamma I = \frac{dR}{dt}$, which makes $\frac{dS}{dt} = -\beta SI$. This leads to when $\beta SI - \gamma I > 0$, the number of infections increases, and $\beta SI > \gamma I$, dividing each term with *I*, and then γ, we have the following form: $(\frac{\beta S}{\gamma} > 1) \equiv (R_0 > 1)$, which shows that $R_0 \approx \frac{\beta S}{\gamma}$. Let $\tilde{S}, \tilde{I}, \tilde{R} \in \mathbb{R}$ be the percentage of overall population such that $\tilde{S} = \frac{S}{N}, \tilde{I} = \frac{I}{N}, \tilde{R} = \frac{R}{N}$ and $\tilde{S} + \tilde{I} + \tilde{R} = 1$. Note that the $\tilde{S} = 1$ during the initial phases. When integrating the three derivative terms, we can measure the differences affecting the overall population throughout the time. We observe the two cases when the proportion of the patient zero $\in \tilde{I}$ is 0.0001 (0.01%, case 1) and 0.01 (1%, case 2). The observation after multiple simulation offers that the two of the time frames t and t' where $\frac{d\tilde{I}}{dt} = 0$ in case 1 and t' where $\frac{d\tilde{I}}{dt'} = 0$ has the relationship of $t > t'$. However, the results have shown that $\sum_{\forall t} R_t \approx \sum_{\forall t'} R_{t'}$ where R_t notes the number of recovered at the time *t*. To be specific, $|\sum_{\forall t} R_t| - |\sum_{\forall t'} R_{t'}| \approx 0.1$, which does not hold relatively large difference compared to the significant gap between the two inputs.

In order to obtain further explainability, merging additional terms based on the SIR model would increase the credibility of the primary model. For instance, adding a number of exposed (E) into the model makes it an SEIR model where E indicates that the host is indeed infected, but it does not transmit the virus to others for a certain preliminary period. Also, multiple terms were added, such as the SEIQRS model which Q indicates the number of Quarantine and the last letter S is identical to the initial S (Susceptible). This implies that the patient who survived the infected set loses the immune system or does not produce the immune system, which directs the patient into set S. These settings are considered typical in the real-world scenario since the COVID-19 era is being elongated, and new variant virus types are emerging every day. The variations of the terms from this expanded model incorporate significant and distinct phases compared to the original SIR model. The notable difference is that it possesses the waves of the surge in infection cases, such that stochastic inference implies reinfection. Thus, new waves of explosive infection occur repeatedly up to n waves (*i.e.*, $n = 3$ in the SEIQRS model).

Fundamentally, accurate prediction of the target variables is an arduous task since the variation of variables is a dynamic process that is constantly being changed. For example, variables β, γ are directly being affected by the current preventive measures and other factors that might alter the values of k in terms of the R and R_0. Types of research were conducted to enhance the predictability of the metrics. Bubar et al. [12] suggested five vaccination distribution schemes analyzed based on the standard of age stratification and showed the coinciding result efficacy after vaccination simulations with SEIR [21] (Susceptible, Exposed, Infectious, Recovered) simulation model. Our work is similar to what this manuscript presents, concerning the derivatives throughout the time of major factors such as infection, death, vaccinated, and recovered cases. Our simulation

considers the various properties of the subjects, and more experimental cases were reflected compared to [21] to organize a realistic simulation framework.

Nguemdjo et al. [32] suggested the R_0 index based on the COVID-19 patient during the March-April 2020 in the city of Cameroon. It utilized the SIR model to observe the constant c, and with different input values, the authors have empirically suggested the diverse outcome and their descriptive statistics of the infection cases result. Similarly, Ajbar et al. [33, 34] applied the classic SIR model but merged it with other factors of non-linear removal rates and effects of media on public awareness based on the COVID-19 patient data of Saudi Arabia. The authors conducted an analysis of computing the statistics implementing the Hopf bifurcations of upper and lower cycles. Another interesting research was proposed (Alanazi et al., 2020) [35] that simulates forecasting the three COVID-19 scenarios: No actions, Lockdown, and new medicine, and these categories are fundamental scenarios that help to measure the baselines of ongoing pandemics. They suggested the new SIR-F model, where F indicates the 'fatal with the confirmation' based on the time-series forecasting machine learning algorithm. Chen et al. [36] numerically analyzed the overall progress of spreading COVID-19 in the affiliated community with an adaptive time-dependent SIR model and the period of turning the reproduction number less than 1 with the COVID-19 dataset of China. Moreover, the authors attempted to explain the impact of social distancing on asymptomatic COVID-19 patients, adding two factors: detectable infection and undetectable cases. By the same token, SIR models calibrating the time-dependent parameters based on the deep learning model were proposed (Jo et al., 2020) [37], validating their model dynamics with the South Korea COVID-19 dataset. Likewise, diverse enhanced SIR models for COVID-19 were suggested, and Ram et al. [38] devised a modified age-structured SIR model implementing the COVID-19 dataset from the state of Washington in the US, including their

affiliated counties. Through the county level age-stratified analysis of multiple perspectives such as social distancing, population statistics, and the operation of public service (school, work) policies. On the contrary, Moein et al. [39] proposed their work, arguing the low performance of the SIR model with the dataset of Iran, Isfahan province, during Feb. 14th to April 11th, raising refutations of the current research about predicting the overall phase of COVID-19 transmission progress.

Based on the SIR model analysis, this research adds another metric of vaccination, and we validate the proliferation of the immunization through the simulation on the ABM. Our research is the first research to construct the ABM that is analogous to real-world and simulate the vaccination to observe the statistical impact that offers dynamics of COVID-19 dispersion. Through thorough COVID-19 patient analysis concerning clinical features and social features, we compute the potential risk (EVI) that quantifies the internal (mortality) and external (infection rate) risk levels. Through ABM-based simulation, we estimate the vaccination effect among the given randomly convoluted ABM-model benchmark dataset and compare the impact of essential metrics between other vaccination criteria: infection cases, death cases, and death rate. The validation standard offers a numerical comparison of existing metrics and demonstrates the effectiveness with respect to propagating the vaccines into the optimal subjects concerning both infection rate and death rate regarding its internal and external factors.

3 MORTALITY AND CLINICAL FACTORS

The death rate of the infected host has been shown to involve a high impact concerning internal conditions. In a general perspective, fatality, when contracted by an infectious disease such as COVID-19, corresponds with the pertaining intrinsic health issues. In a similar context, the correlation between the lethality of the disease and the level of health is shown to have a positive correlation, which can be interpreted as the risk accumulates, depending on the internal basis. In addition, the immune system that produces the antibody when the virus is injected differs from the independent subject, and to constantly generate the profitable cells, fully activating the immune system is imperative. This section concentrates on the statistics of the three fundamental and influential factors concerning the death rate: age-stratified groups, types of comorbidities, and gender. Based on the COVID-19 benchmark dataset with those factors, our aim is to estimate the internal risk by statistically measuring the susceptibility.

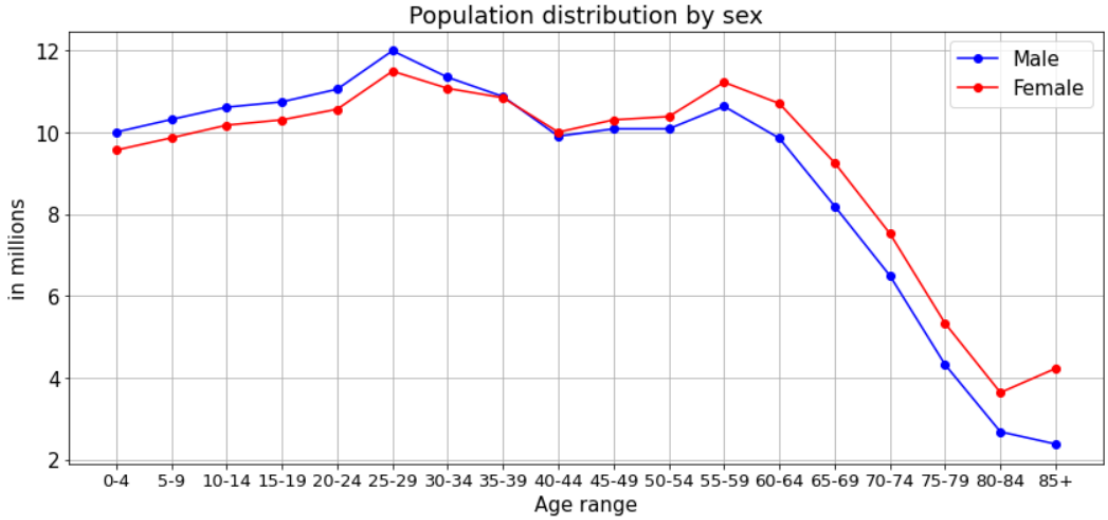
3.1 Mortality of Age-stratified Groups and Gender

The existing medical institutions, such as CDC and WHO, have announced that there is a clear sign that age and gender are vital properties to measure the subject's risk [40, 41]. Most patients that show acute symptoms tend to possess relatively higher ages than patients with a lower fatality, which causes a higher level of death rate. Among the rate indexes that were posted by the Worldometer, 2021; Case Fatality Rate (CFR), Infection Fatality Rate (IFR), Crude mortality Rate (CMR) that was computed based on the age groups, the result has shown that the groups with more significant age level are susceptible to the disease with high mortality rate. In a sense, the age group does offer a precise risk level, as the most disease does. Furthermore, from the Global Health 50/50, most of the countries with open COVID-19 patient datasets had shown to have a higher

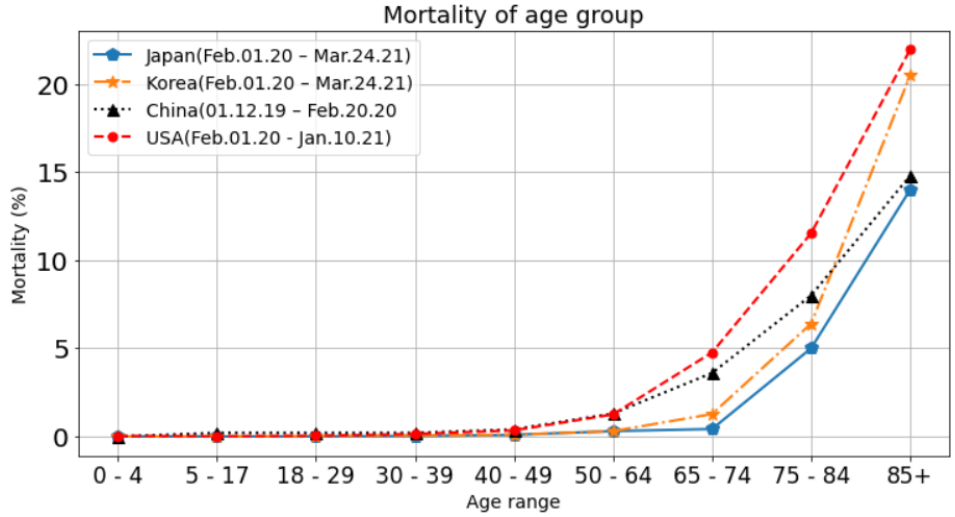
death rate for male patients instead for female patients. Several nations have shown this in drastic statistics. For example, until 2020 December, Thailand, Nigeria, Yemen, Bangladesh, Afghanistan, Pakistan, and Malawi had shown that almost 75% of the death cases have occurred from male patients. India and Mexico had around 64% death cases for males and 58% for Brazil. There are diverse perspectives assert a valid explanation related to this statistical observation. First, many scientists suggest that females possess a relatively more robust immune system. Males are more frequently born (105:100); however, a year after the infants are born, the ratio approximates to equivalent proportion in worldwide statistics. Also, it is undoubtedly true that the lifespan of females much higher than males (*i.e.*, around 6~8 years), and the ratio of gender population after the age of 100, females are four times larger than males. This phenomenon was validated in another pandemic SARS in 2003 and during the Spanish flu. There are assertions that this originates from daily habits, such as the male ratio of smokers being significantly large or males tend to involve in hazardous affairs. Also, females possess two X chromosomes, and this functions as an advantage for performing better tasks when one X chromosome is vulnerable; the other entity can operate as a substitution, whereas having a single X chromosome mainly illustrates a clear sign of weakness. Moreover, hormones that secrete primarily from the male, such as testosterone, tend to debilitate the immune system.

Let s_i be an individual subject, with $S = \bigcup_{\forall i} s_i$ where S refers to the population. We set notation of the mortality of subject s_i as $\mathbb{M}(\cdot)$, and $\mathbb{M}(s_{\exists i}) \coloneqq r/S$ where the r indicates the number of death cases. As the statistical data presents**,** subjects that are affiliated with relatively higher age groups have greater mortality, as shown in Figure 3.1 (b), in four countries during the given period. Let age group $A \coloneqq \{a_k | 1 \leq k \leq 9, k \in \mathbb{N}\}$, where a_k denotes the number of subjects among the age range of 0-4, 5-17, 18-29, 30-39, 40-49, 50-64, 65-74, 75-84, 85+

respectively. We denote that $a_{\exists k} = \bigcup_{\forall i'} s_{(k,i')}$, and $s_{(k,i')}$ refers to a subject with the age range of corresponding a_k where i' indicates the index. Currently, $n(S) = n(A) = \sum_{\forall k} n(a_k)$ and $\mathbb{M}\left(s_{(k,i)}\right) := r_k / a_k$.



(a)



(b)

Figure 3.1 (a) COVID-19 patients in USA assorted by sex (b) COVID-19 Mortality by given age-stratified groups in corresponding countries by a given period

Based on the collected raw COVID-19 patient dataset [42, 43], we generalize the required parameters by processing the data, and the mechanisms are as follows. As of March 31, 2021, the accumulated proportion of COVID-19 cases in gender (male and female) is 6,277,679 (Female; I_F) and 5,750,585 (male, I_M). The relationship of those values shows that $I_F = I_M \cdot 1.09$, and $I_F : I_M = 52.2 : 47.8$, having approximately 4.4% higher for the I_F as other studies have specified such as [40]. Let the conditional probability $P(r_G|I) := \mu(s_G)$ such that $I_F \cup I_M = I, G \in sex = \{F, M\}$. The input s_G indicates the subject that has the corresponding sex G and r_G refers to the number of death cases in assigned sex types, with the ratio of $\mathbb{M}(s_F) : \mathbb{M}(s_M) = 0.559 : 0.441$. Proportionally, both terms differ by the values of 0.5 ± 0.059, where we denote $\pm 0.059 = P_G$ as shown in equation (1). The indicating subject's death rate that was allocated with the age index *k* and sex *G* can be defined as $\mathbb{M}(s_G) \cdot (1 + P_G) := \mathbb{M}(s_{(k,G)})$, followed by normalizing the component with min-max normalization, assigning the values $\mathbb{M}(s_{(k,G)}) \in [0,1]$. Although the min-max normalization computation is known to be less productive when indicating the relationship if the dataset incorporates outliers, however, our case is suitable to use it for no outliers exist since *G* is deterministic. Equation (2) defines the statistical risk of $\mathbb{M}(s_{(k,G)})$ as follows.

$$P_G := \frac{\mathbb{M}(s_G)}{\sum_{G \in sex} s_G} - 0.5 \tag{1}$$

$$\mathbb{M}(s_{(k,G)}) := \frac{\mathbb{M}(s_k)(P_G + 1) - min\,(\mathbb{M}(s_k)(P_G + 1))}{max(\mathbb{M}(s_k)(P_G + 1)) - min\,(\mathbb{M}(s_k)(P_G + 1))} \tag{2}$$

3.2 Comorbidities and Mortality

If the patient possesses Comorbidities, the death rate is relatively much higher than the patient with no underlying disease [44-48]. Past COVID-19 studies have observed this phenomenon, and researchers have asserted that there is a strong correlation between the death rate and several underlying diseases, which implies that comorbidities escalate the potential risk. In [44], the research shows that up to 90% of hospitalized COVID-19 patients have suffered from comorbidities before the infection. For example, according to the Korea Disease Control and Prevention Agency, COVID-19 patient death cases (5,382, Jan. 2022) accompanied underlying diseases in a total of 10,366, having 1.9 comorbidities per patient. Among 5,382 cases, only 169 cases did not have any comorbidities. Depending on the country, the comorbidity types that lethally effects the COVID-19 patients are observed as respiratory disease (*i.e.*, chronic lung problems, asthma) that directly correspond to the COVID-19 (respiratory disease) impact, and circulatory system disease (*i.e.*, heart disease, artery disease), which prevents building the immune system. In general, patients who suffered from diseases, especially respiratory disease, cardiac disease, and cardiovascular disease, were found to have relatively higher death rates from the previous statistics. Likewise, the comorbidity of the patient provides diverse information and evidence of future prognosis when estimating the risk. The COVID-19 patient dataset [42] used in this research categorizes the types of illness into 22 large-scale categories established in WHO's International Statistical Classification of Disease and Related Health Problems (ICD, 2020) [49], which was shown in Table 3.1. Moreover, the disease types can be further be classified into more detailed disease types via ICD criteria.

$\mathbb{M}\left(s_{(k,G,D)}\right)$ indicates the subject's mortality with three internal attributes of *k*, *G*, and *D*. Let $D \coloneqq \bigcup_{i=0}^{22} \mathbb{C}_i$, and through $\exists \mathbb{C}_i$, we measure the Pearson correlation coefficient (PCC) among

given age group as shown in Figure 3.2. PCC is expressed with $\rho_{(k,D)} := \frac{Cov(\mathbb{M}_{(s,k)}, \mathbb{M}_{(s,k)})}{\sigma(\mathbb{M}(s_D))}$. Mostly, it shows $\rho_{(k,D'')} \geq 0.8$ where $\sigma(q)$ expresses the standard deviation of input list q, and $D'' := D - \{\mathbb{C}_3, \mathbb{C}_{14}, \mathbb{C}_{15}, \mathbb{C}_{16}, \mathbb{C}_{17}, \mathbb{C}_{18}, \mathbb{C}_{19}\}$. For most of $\rho_{(k,D')} \geq 0.7$, where $D' := D'' + \{\mathbb{C}_3, \mathbb{C}_{14}, \mathbb{C}_{15}, \mathbb{C}_{16}\}$. This diversity is due to the category of disease shows slightly different statistics among the age groups. Representatively, $\mathbb{C}_{17}$ indicates the group of obesity patients with the age-stratified group of 55~64 had the highest number of death cases. On the contrary, $\mathbb{C}_{18}$(Alzheimer) and $\mathbb{C}_{19}$(Dementia) show to have an exponential relationship as the age increments linearly. Apart from those features, most of the $\mathbb{M}(s_D)$ shows a positive correlation relationship with $\mathbb{M}(s_k)$ when aligned by age. Until the age group of 55~64, the $\Delta s_{(k,k+1)}$ tend to have an incremental phase where $\Delta s_{(k,k+1,D')} = \left|\mathbb{M}\left(s_{(k,D')}\right) - \mathbb{M}\left(s_{k+1,D'}\right)\right|$. After $\Delta s_{(k,k+1,D')} = 0$, the $\Delta s_{(k,k+1)} < 0$ starts to decrease, having $\Delta s_{(\tilde{k},\tilde{k}+1,D')} > \Delta s_{(\bar{k},\bar{k}+1,D')}$ where $\bar{k} := index > 65$, and $\tilde{k} := index \leq 65$. In Figure 3.3, it displays that the linear correlation $PCC(\mathbb{M}_k(s_D))$ (*i.e.*, $\mathbb{M}(s_D)$ sorted by the order of k), which implicates that $\mu(\mathbb{M}_{s_{(D,\exists k)}})$ and $\mu(\mathbb{M}_{s_{(D,\exists k+L)}})$ where $\mu(q)$ denotes the average of set of inputs of q and $(L \neq 0) \in \mathbb{R}$, $1 \leq (k + L) \neq k \leq 9$ gradually inclines to be distinct as $|L_k| \gg |L_{k-c}|$ and $k \gg c \in \mathbb{N}$. Table 3.1 displays the COVID-19 death cases sorted by the corresponding comorbidity by age groups. Note that the comorbidity types indicated in Table 3.1 are categorized in Table 3.2. Equation (3) shows the mortality with respect to k, D: (*i.e.*, $\beta_{s_{(k,D)}}$) based on the statistics of Table 3.1.

$$\beta(s_{(k,D)}) := \frac{\mathbb{M}\left(s_{(k,D)}\right)^2}{\mathbb{M}(s_k)\mathbb{M}(s_D)} \tag{3}$$

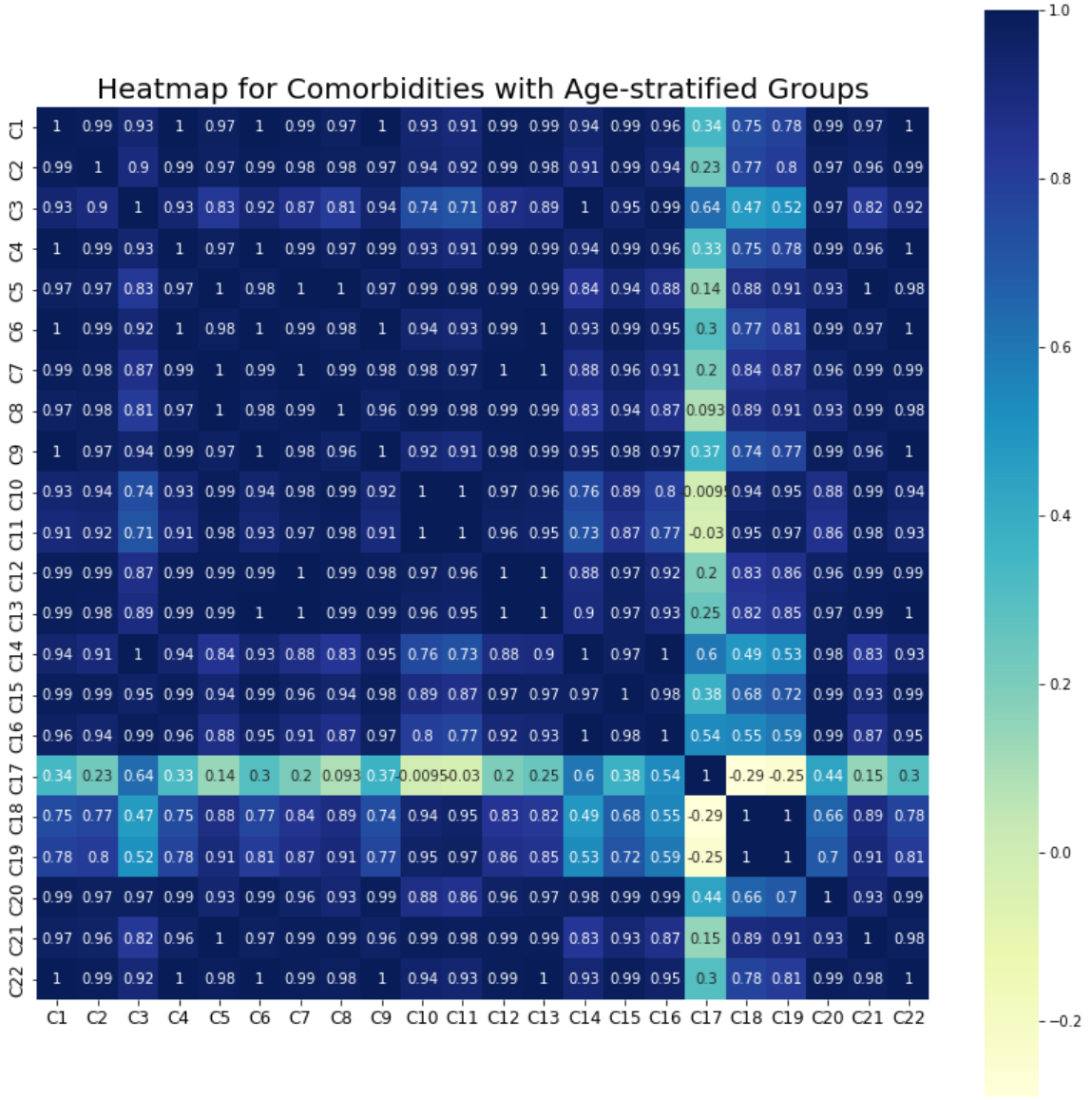


Figure 3.2 Heatmap indicating the correlations between each age group based on $\mathbb{M}(s_k)$

Table 3.1 Comorbidity Types and Corresponding Number of Casualties in each Age-Stratified Groups (Comorbidity Types are explained in Table 3.2)

Comorbidity	0-24	25-34	35-44	45-54	55-64	65-74	75-84	85+
$\mathbb{C}_1$	175	853	2,174	6,220	15,295	25,749	30,258	30,239
$\mathbb{C}_2$	36	88	203	594	2,334	5,577	7,551	6,749

$\mathbb{C}_3$	86	328	895	2,527	5,582	8,367	7,667	5,659
$\mathbb{C}_4$	153	613	1,582	4,726	11,840	21,324	25,811	24,738
$\mathbb{C}_5$	10	38	88	234	580	1,062	1,427	1,837
$\mathbb{C}_6$	25	89	211	530	1,273	2,267	2,717	2,787
$\mathbb{C}_7$	25	163	667	2,261	6,375	11,807	14,966	17,828
$\mathbb{C}_8$	5	38	154	692	2,640	6,085	9,088	10,804
$\mathbb{C}_9$	76	278	678	1,930	4,344	6,869	7,442	7,944
$\mathbb{C}_{10}$	14	36	107	375	1,238	3,107	5,240	7,144
$\mathbb{C}_{11}$	9	54	126	444	1,442	3,113	5,130	7,899
$\mathbb{C}_{12}$	9	41	137	466	1,452	2,885	3,737	4,209
$\mathbb{C}_{13}$	65	160	361	845	1,945	3,380	3,970	4,688
$\mathbb{C}_{14}$	56	223	560	1,697	4,030	6,652	6,177	4,435
$\mathbb{C}_{15}$	39	54	150	438	1,650	3,242	3,711	3,886
$\mathbb{C}_{16}$	64	257	899	2,723	6,714	11,198	11,391	8,427
$\mathbb{C}_{17}$	123	430	960	1,670	2,418	2,446	1,321	358
$\mathbb{C}_{18}$	0	0	0	6	77	670	2,848	6,308
$\mathbb{C}_{19}$	0	1	2	31	413	2,605	8,548	17,068
$\mathbb{C}_{20}$	29	180	463	1,388	3,426	5,891	6,249	5,574
$\mathbb{C}_{21}$	90	262	337	550	1,103	1,759	2,225	3,017
$\mathbb{C}_{22}$	435	1,175	2,631	6,997	17,581	31,390	36,470	38,705

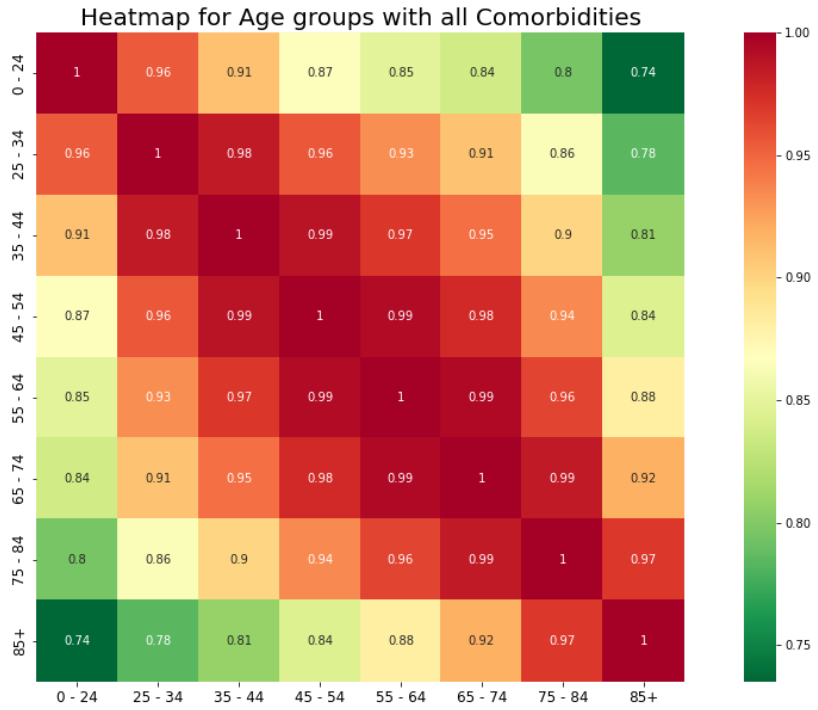


Figure 3.3 Heatmap indicating the correlations between each age group based on $\mathbb{M}(s_D)$

Table 3.2 Comorbidity Types and corresponding Index

Comorbidity Name	Index
Influenza and Pneumonia	$\mathbb{C}_1$
Chronic lower respiratory diseases	$\mathbb{C}_2$
Adult respiratory distress syndrome	$\mathbb{C}_3$
Respiratory failure	$\mathbb{C}_4$
Respiratory arrest	$\mathbb{C}_5$
Other respiratory diseases	$\mathbb{C}_6$
Hypertensive disease	$\mathbb{C}_7$
Ischemic heart disease	$\mathbb{C}_8$
Cardiac arrest	$\mathbb{C}_9$

Cardiac arrhythmia	$\mathbb{C}_{10}$
Heart failure	$\mathbb{C}_{11}$
Cerebrovascular diseases	$\mathbb{C}_{12}$
Other circulatory diseases	$\mathbb{C}_{13}$
Sepsis	$\mathbb{C}_{14}$
Malignant neoplasms	$\mathbb{C}_{15}$
Diabetes	$\mathbb{C}_{16}$
Obesity	$\mathbb{C}_{17}$
Alzheimer disease	$\mathbb{C}_{18}$
Vascular and unspecified dementia	$\mathbb{C}_{19}$
Renal failure	$\mathbb{C}_{20}$
Injury poisoning other events	$\mathbb{C}_{21}$
Other conditions and causes	$\mathbb{C}_{22}$

Although there are evident statistics that comorbidity operates as a fatal factor in aggravating the patient's status, the exact influence regarding the number of comorbidities is not yet clearly suggested. However, we assume that the patients may possess multiple comorbidities and set the maximum number of underlying diseases as three. The exact database for the disorder for the comorbidity types of the COVID-19 patient is not currently publicly opened (*i.e.*, most of the open dataset publicity opened the total number of comorbidities among the total patients). In this thesis, we randomly assigned the number of comorbidities q_i, where $q_i \in [0,3]$ and $q_i \in \mathbb{N}$. We explain the additional comorbidity part regarding how we implemented in our algorithms in detail in section 4.3. After the aggregation, the outcome normalizes the value using the min-max

scaler. Each gathered factor is considered independent and we linearly aggregate the factors to numerically define the risk. The final mortality $\mathbb{M}(s_{(D,G,k)})$ computation sequence is shown in equation (4). $\mathbb{M}(s_{(D,G,k)})$ evaluates the risk that the subject possesses is statistically shown from the past COVID-19 patients. Another meaning that $\mathbb{M}(s_{(D,G,k)})$ connotates is the death rate of the subject with its internal conditions, and this offers the advantage of leveling the risk with given features compared to the other past patients.

$$\mathbb{M}\left(s_{(D,G,k)}\right) \coloneqq \frac{\mathbb{M}\left(s_{(G,k)}\right) + \sum_{\forall q_i} \beta(s_{(D,k)}) - \min\left(\mathbb{M}\left(s_{(G,k)}\right) + \sum_{\forall q_i} \beta(s_{(D,k)})\right)}{max\left(\mathbb{M}\left(s_{(G,k)}\right) + \sum_{\forall q_i} \beta\left(s_{(D,k)}\right)\right) - min\left(\mathbb{M}\left(s_{(G,k)}\right) + \sum_{\forall q_i} \beta(s_{(D,k)})\right.} \tag{4}$$

Machine learning-based approaches that train the classifier model to predict the soft label (probability) of death require a significantly higher computational amount by locating the non-linear decision boundary, also detecting the outliers such as a death, in this case, is likely to produce an overfitting model since a number of cured cases is notably more significant than the death cases, which may not function properly when it comes to actual practice. Finally, the purpose of the machine learning model is to make classification, whereas our objective is to observe the level of the risk by aggregating multiple features. Thus, we linearly aggregated the given risks on each feature independently, which explicitly reflects the death rates of each given feature.

4 INFECTION RATE AND NETWORK CENTRALITY

In this section, we cover the infection rate, which is another factor that we must downsize. When predicting the virus proliferation dynamics, computing the infective trajectory and its probability of infection through the routes are major objectives. Diverse epidemiological research was conducted to define the computation model, mathematically approximating the authentic dispersion trend. Contrary to the death rate, external factors such as the environments trigger the correlation of the infection rate. In this thesis, we assume that when the virus is being transmitted from the patient to the non-infected, the internal immune system effect of contracting the disease is neglected. After the virus infects the host, it is a deterministic process with clear vital signs to be observed. However, releasing the information about the infection without any symptoms or specific measures to clarify whether the virus was inside the host is a complex matter. Therefore, we only focus on the deterministic virus injection and regulate the infection in a stochastical fashion. According to the R index in section 2.3, metric c differs in the effect of the infection-preventative measures such as quarantine, social distancing, and encouraging hygienic activities (*e.g.*, wearing masks, washing hands). Apart from directive schemes, inherent characteristics such as depending the regional features, population density, local lifestyle, age variance, GDP per capita, etc., tend to affect the spread indirectly.

The infection dynamics are deemed a fundamentally convoluted process, consisting of diverse dependent and independent variables. Various researchers tend to adopt datasets under the assumption with a significant level of statistics in regional boundaries, locating environmental hyperparameters. In this research, our interest focuses on discovering the impact of the host concerning their social network. Intuitively, when the person who holds an impactive role with relatively more extensive physical contact among its associated community gets infected, its

impact will likely be significantly more extensive than that of an ordinary person. In other words, we lead to the question of the quantitative difference between super spreaders and standard spreaders with respect to the infection ratio among the overall community. To track down the answer to this question, we implement the vaccine propagation simulation in the ABM to infer the variations of three target metrics and validate the performance by comparing it with the other vaccination sequences' metric ratio.

4.1 Agent-Based Model with Graph-Network Structure

Essentially, a virus is diffused through physical contact from the infected host to another. In our research, we assume that the virus is infected under the constraints of proximate distance, similar to COVID-19. Since infection is one-way propagation perpetrated from the infected to the non-infected, it is feasible to define this relationship through graph network structure [50]. Through distributed environments with numerous potential hosts and each host with various types of features, the physical contacts are dynamically and perpetually occurring in real-time. This situation can be modeled via graph networks, with dissemination models such as probabilistic model (random walk, stochastic model) or deterministic model.

Let $g = (V, E)$ where V is a set of vertices, and E is a set consisted of edges where $V = \{v_i | 1 \leq i \leq N, (i, N) \in \mathbb{N}\}$, N refers to the total number of agents (nodes) that is a subset of g and $E = \{< v_a, v_b > | (v_a, v_b) \in V \times V, 1 \leq (a, b) \in \mathbb{N} \times \mathbb{N} \leq T \in \mathbb{N}\}$. g is a finite graph, and $< v_a, v_b >$ denotes the directed edges with source node v_a and sink node v_b. Recall that the infection transmission is a one-way directive process, with a source and destination where infected agents transmit the virus to the physically connected non-infected. Thus, we select the directed graph structure, and let $E_a = \bigcup_{\forall b} < v_a, v_b >$. $n(E_a)$ is the number of edges which the initial node is v_a,

and it is connected to $\exists v_b$ where b specifies the index of i, the neighbor node of the v_a. If v_a that has $\max n(E_{\forall a})$ is infected from the neighbor v_b such that $v_a = True$, the $\frac{dE_a}{dt} > 0$; this escalates the diffusion pace with a high probability due to the significant centrality value among the affiliated community compared to other nodes with relatively lower E_a.

Each agent in the associated ABM occupies a certain portion of importance [50], and we adopt the five network centrality schemes to evaluate the importance or possible degree of impact that the target node is capable of invoking. The five centralities are as follows: Degree centrality [52], Closeness centrality [51], Betweenness centrality [53], Eigenvector centrality [54], and PageRank [55]. Each brief computation schemes are as follows. Degree centrality evaluates its index through the number of degrees; $n(E_a)$. Closeness centrality normalizes the standard distance to each connected node in the shortest path, which follows the basis of 'important node is proximate to other nodes'. In mathematical values, we have $\frac{n(V)-1}{\sum_{a\neq b}^{n(V)} E_a}$, which is an equation of the normalized Closeness centrality. Betweenness centrality focuses on the paths that a source node visits in order to reach the sink node. Then, it linearly accumulates the number of probabilities of visiting a particular node. To express Betweenness centrality (*BC*) into mathematical form, we have centrality of node c as follows: $BC(v_c) = \sum_{v_a \neq v_b \neq v_c, a>b} \frac{\varepsilon_{(a,b)}(v_a)}{\varepsilon_{(a,b)}}$, where $\varepsilon_{(a,b)}$ indicates the number of shortest paths between the node a and b, and $\varepsilon_{(a,b)}(v_a)$ denotes the number of shortest paths between the node a and b that passes node v_a. Eigenvector Centrality (*CE*) implies that the importance of the target node is set by the significance level of the neighbors, indicating that if such a target node possesses many popular nodes, then this target is also important. It computes the largest eigenvalue and corresponding eigenvector, having $CE(v_a) \propto \sum_{v_b} A_{(a,b)} CE(v_b)$ where $A_{(\exists a, \exists b)}$ is an adjacency matrix with existing nodes a, b. The PageRank score is an upgraded

version of the Eigenvector centrality. Thus, Eigenvector centrality and PageRank share a similar high-level idea: it evaluates the connected neighbor nodes without directly assessing the number of interconnections. We explain PageRank in detail in section 4.2. In the collection of five centrality metrics, we take the worst-case scenario, a centrality that propagates the virus within the shortest amount of time unit when the equivalent ABM setting was given.

4.2 Selecting Optimal Centrality

Among multiple centrality criteria, our objective is to select the metric shown to have the worst effect in our proliferation scenario. In order to choose the worst-case centrality, we measure and compare the amount of period that takes to be infected from the initial spreaders until the edges reach the leaf node, which halts the dissemination. To start, we build the three ABM with $n(V) = 1000,\ 5000, 10000$. By using the physical contact per day dataset collected and suggested by the Del Valle et al. [26], we set the $n(E)$ for each node based on the feature of the node and corresponding value in [26]. Recall that we assumed that the infection spreads through close contact exposure. The number of edges was allocated based on the ratio of US population demographics by age-stratified groups in 2020 [56]. Each age group has a designated contact frequency, based on the statistics shown in [26], which is shown in Table 4.1. Between the range of $1 \leq n(E_a) \leq 2 \cdot RCF$, we randomly assign the $n(E_a)$. RCF (Round-up Contact Frequency) denotes the frequency values for the coinciding age group that was rounded up to an integer. for the coinciding age group that was rounded up to an integer. This simulation is to select the worst-case centrality that has the fastest dispersion velocity, and we use a 100% infection rate as a default setting when transmitting the virus. In the simulation, nodes $v_{\exists a}$ that are connected to the infected nodes $v_{\exists b}$ are being contaminated in each time step *t*; where we denote this as $< v_a, v_b >_t$. We

visualize the diffusion in Figure 4.1, and display the pseudocode of the spreading algorithm in section 5.1.

Table 4.1 Population Ratio and Corresponding Contact Frequency

Age	Population Ratio	Contact Frequency	RCF
0-4	0.068	10.216	10
5-9	0.061	14.812	15
10-14	0.063	18.224	18
15-19	0.064	17.582	18
20-29	0.137	13.573	14
30-39	0.135	14.142	14
40-49	0.123	13.830	14
50-59	0.129	12.308	12
60-69	0.116	9.216	9
70+	0.112	6.898	7

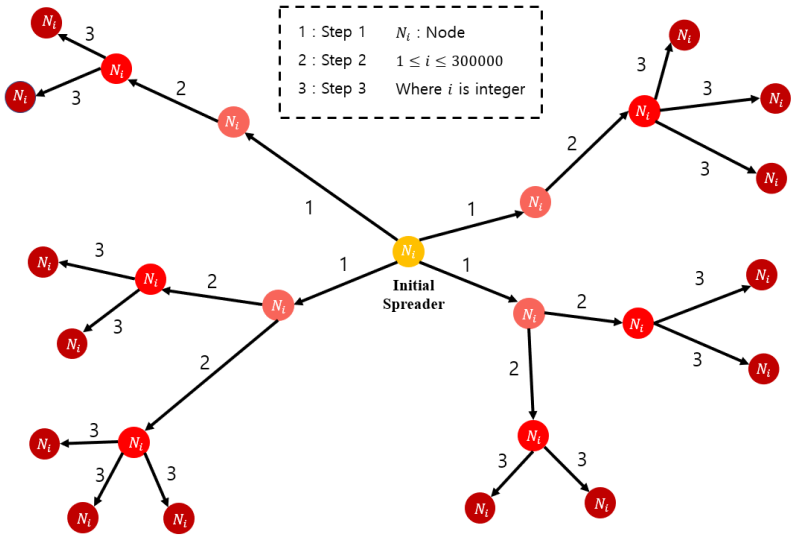


Figure 4.1 Time steps t (=1,2,3) of spreading disease in given Graph Network

A single edge can be taken from the source node during a one-time step unit, reaching other connected nodes. This random walk-based algorithm is designed via recursion and depth-first-search. For the pseudocode, authors can refer to algorithm 1 in section 5.1 for more details. When the initial nodes have significant centrality, it leads to a faster spread. The initial ten spreaders are $\tilde{\rho} = \bigcup_{10} \tilde{v}_i$ where $\tilde{v}_i = \max(\bigcup_{\forall i} \phi(v_i))$, and $\phi(v_i)$ denotes the centrality of the input v_i and $\tilde{v}_i$ is sampled without replacement. We experimentally validate the selection of $\max(\bigcup_{\forall i} \phi(v_i))$ with comparing $\min(\bigcup_{\forall i} \phi(v_i))$ in the next section. After 50 trials of each virus dispersion when setting the ten initial patients with the maximum degree of centrality metrics on three different numbers of ABMs, the results are shown in Table 4.2. Values indicated in Table 4.2 are the average of 50 trials. Each value is the average value of 50 trials (updated to a new graph structure of each trial), and the number inside the parenthesis illustrates the standard deviation. Each attempt was conducted with the newly generated random graph structure, $g_i \neq g_{i'}$, where $i \neq i'$. The results show that initial patients with the highest PageRank spread the virus within the shortest amount of time unit, which verifies the worst-case scenario. Thus, we select the PageRank scheme for measuring the infection rate and compute the EVI. PageRank was suggested by the founders of Google, and it is widely known as the mechanism for Google's web search engine [55]. It ranks and judges the importance of the node in the graph structure by computing the quantity of the edges also the quality of the connected neighbor nodes. The original format of the PageRank is shown in equation (5), where $PR(\cdot)$ refers to the PageRank value, and we set the damping factor (q) to 0.85.

$$PR(v_i) = \frac{1-q}{N} + q \sum_i \frac{PR(v_i)}{NumLinks(v_i)} \tag{5}$$

Table 4.2 Average Period Duration among given Centrality

Centrality	Number of Nodes			Average
	1,000	5,000	10,000	
Degree	271.8 (±4.2)	331.4 (±3.3)	332.3 (±2.9)	311.8 (±3.5)
Closeness	129.5 (±20.7)	145.5 (±22.1)	143.2 (±12.0)	139.4 (±18.3)
Betweenness	129.5 (±20.7)	145.5 (±22.1)	143.2 (±12.0)	139.4 (±18.3)
Eigenvector	130.4 (±19.9)	144.4 (±31.1)	156.4 (±11.2)	143.8 (±20.7)
PageRank	116.9 (±19.4)	139.4 (±14.3)	103.7 (±10.7)	123.1 (±14.8)

4.3 Simulation construction and EVI

To conduct the empirical simulation for the pandemic propagation procedure, this section constructs the ABM and computes the final EVI. The descending order of PageRank efficiently suppresses the contamination speed. We assemble the calculated factors to evaluate the final EVI. Similar to the three different agent numbers of ABM, we use 300,000 agents, appointing unique features to each agent based on the statistical ratio [42, 43]. For example, let the male population of age group 35~44% take p%, and by assigning $n(V) \cdot p \cdot 0.01$ nodes with a gender of male, and random age between 35~44. The distribution of a set that is composed of similar features has a uniform distribution. Also, the allocation of underlying disease types with a maximum number of three utilizes the dataset (*i.e.*, death cases of the COVID-19 patient database) in section 3.2. By calculating the death rate among the infected population, we have 1.45% of mortality. However, we intentionally increase this value to 7.25% (1.45×5) because this value triggers sparsity among the ABM, and comprehensible variations can be observed after the variation ends. Likewise, comorbidity types were selected for the agents with optimal statistical proportions. Recall that the

number of comorbidities was randomly chosen between 1~3 since a patient may have multiple underlying diseases. Graph edges are given with the format of section 4.2., based on the physical contact by age group statistics. Furthermore, the graph structure was subdivided to incorporate geographical constraints. By segregating the graph into five large clusters: $\tilde{C}_{1\leq u\leq 5}$, and each $\exists C_u$ is subdivided again into six groups $\tilde{c}_{(\exists u,i)} \in \tilde{C}_{\exists u}$ where $1 \leq i \leq 6$ (refer to Figure 4.2 for visualization)**.** This partition is reasonable since the virus is initially dispersed in regional boundaries since physical contacts occur only at a close distance within the spatial constraints, and gradually it moves on to other regions. Also, note that each region consists of a statistical ratio of incorporated features such as age groups, in order to accommodate the diversity of regional properties. In addition, $\tilde{c}_{(\exists u,i)}$ are intermittently connected by 20% of the original edges. Figure 4.2 displays the general visualization that shows the interaction (edges) inside the dense community cluster and the connection between $\exists c_{(u,i)}$, as well as $\exists \tilde{C}_u$. The edge density becomes sparser from $\tilde{c}_{(u,i)}$ to $\tilde{C}_u$, which denotes comparatively low physical interaction frequency.

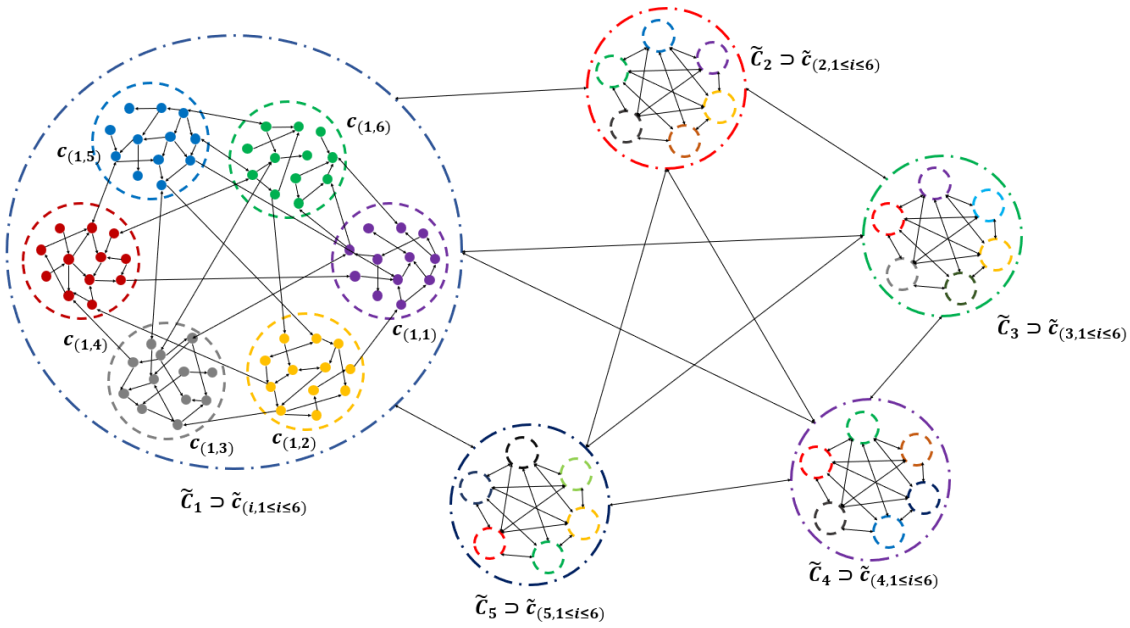


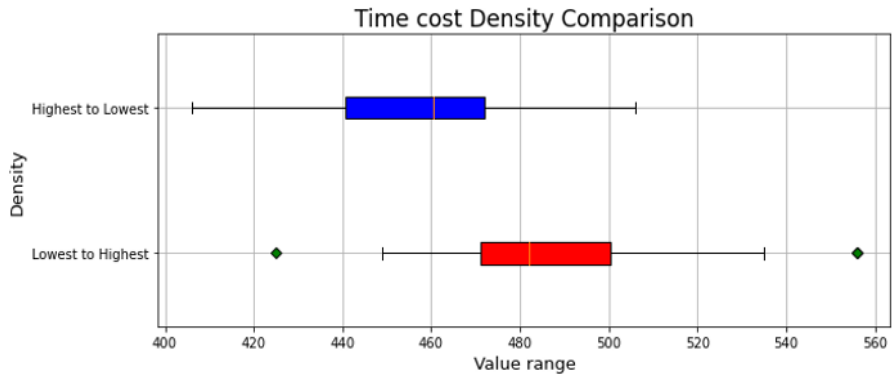
Figure 4.2 The constructed graph network structure for future simulations

Our experiment compares the two cases. In the first case, we select the $\bar{\rho} = \bigcup_{20} v_i$, whereas the second case selects the $\tilde{\rho} = \bigcup_{20} \tilde{v}_i$. This $\bar{\rho}, \tilde{\rho}$ refers to initial spreaders where $\bar{v}_i = \min\left(\bigcup_{\forall i} \phi(v_i)\right)$ and $\tilde{v}_i = \max\left(\bigcup_{\forall i} \phi(v_i)\right)$ with $\phi(v_i)$ denotes PageRank of v_i. The infected node iteratively contaminates other connected nodes, and we measure the cumulative period until the status of the graph does not change. When case $\tilde{\rho}$, the mean of 100 diffusion periods was 456.607-time steps and 485.642-time steps for $\bar{\rho}$. The density distribution of the two cases is displayed in Figure 4.3. This entails that the descending order of diffusion ($\tilde{v}_i$) will lead to faster stabilization, validating the advantage of vaccination based on $\tilde{v}_i$. Equation (6) shows the EVI is computed individually through aggregating the independent $\exists \mathbb{M}(s_{(D,G,k)})$ and $\exists \phi(v_i)$, normalizing factors with the standard normalization and we scale the accumulation with the min-max scaler.

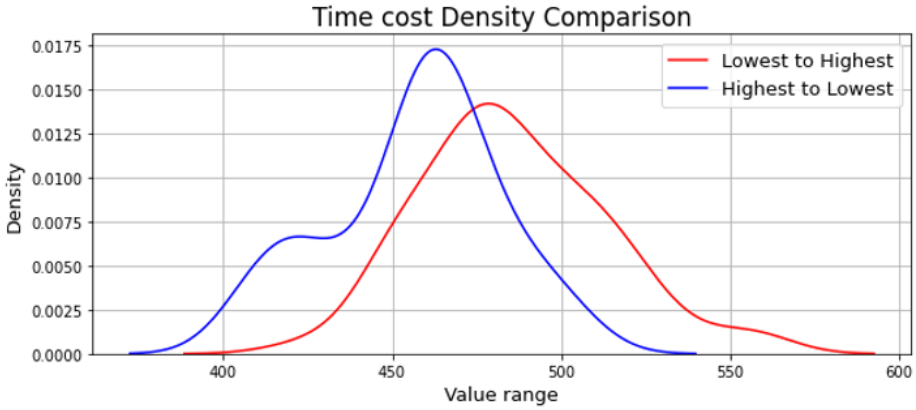
In equation (6), the EVI aggregates both death rate and infection rate for harmonical minimization by calibrating the weights to each term with $W_1 : W_2$ portions. Each weight determines the degree of influence of each factor by adjusting the values where $W_1 + W_2 = 1$; such that $1 - W_1 = W_2$, and $(W_1, W_2) > 0$.

$$EVI = \frac{W_1 A + W_2 B - \min\left(W_1 A + W_2 B\right)}{\max(W_1 A + W_2 B) - \min\left(W_1 A + W_2 B\right)}$$

$$s.t. \quad A = \frac{\mathbb{M}\left(s_{(D,G,k)}\right) - \mu\left(\mathbb{M}\left(s_{G,k}\right) + \sum_{\forall q_i} \beta\left(s_{(D,k)}\right)\right)}{\sigma\left(\mathbb{M}\left(s_{(G,k)}\right) + \sum_{\forall q} \beta\left(s_{(D,k)}\right)\right)} \; and \; B = \frac{\phi(v_i) - \mu(\phi(V))}{\sigma(\phi(V))} \tag{6}$$



(a)



(b)

Figure 4.3 Distribution results of period visualization using box plot (a) and density plot (b)

Every agent in the ABM possesses their individual EVI computed unique personal properties, which can be expressed through the matrix form. For the number of agents in the ABM is 300,000 ($n = 300{,}000$), we build $M_1 = n \times n$ matrix form with row i and column j. The edges are $n_{(i \neq j, j \neq i)} = e_{(i,j)} = 1$ or 0, and the diagonal components $n_{(i=j, j=i)} = \text{EVI}(n_{(i,j)})$. Note that $e_{(i,j)}$ may be equivalent or different since the graph structure is composed of directed edges. Simplex matrix form can be denoted as equation (7), and in section 5.1, we use this matrix basic form to define relationships and extract knowledge.

$$M_1 = \begin{bmatrix} EVI(n_{(i,j)}) & \cdots & e_{(i,j)} \\ \vdots & \ddots & \vdots \\ e_{(i,j)} & \cdots & EVI(n_{(i,j)}) \end{bmatrix} \text{ s. t. } e_{(i,j)} \begin{cases} 1 \;\; if < v_i, v_j >= True \\ 0 \;\; if < v_i, v_j >= False \end{cases}$$

(7)

Since EVI is a linear aggregation, simplex matrix computation is feasible by metricizing the independent multivariate values by combining them into the matrix form. Based on the (6), the matrix form of EVI such as the following equation (8). Through matrix calculation, deriving EVI values can be obtained in an efficient manner. Section 5.1 further utilizes this matrix form to update the virus proliferation status among the ABM.

$$M(EVI) = \frac{\{M(W_1)M(A) + M(W_2)M(A) - \min(M(W_1)M(A) + M(W_2)M(B))\}}{\{\max(M(W_1)M(A) + M(W_2)M(B)) - \min(M(W_1)M(A) + M(W_2)M(B))\}}$$

(8)

Figure 4.4 shows the correlation values of combinations of each feature. The labels inside Figure F1~F11 are as follows: {*population, population density, land area, infection cases, death cases, infection rate, death rate, GDP per capita, Median age, race variance, age variance*} with regards to 50 state-level open datasets of United States. Since our primary concern is to track the characteristics of the death rate and the infection rate, we explore the relationship between those two major properties. The linear correlation coefficient between the infection rate and death rate indicates a value of -0.3. The existing studies have inferred a current trade-off between the infection rate and mortality among the identical population [57-60]. Among the diverse perspectives that explained the causality of this phenomenon, one of the most reliable interpretations is the relation of the age groups. The older generations are mostly the death victims, increasing the death rate. Whereas the younger generations are easily infected, occupying a high infection rate. This shows the trade-off between the generations, with a negative correlation value. Especially when it comes to the problem of which candidate we should inject first leaves us the

dilemma of decreasing the infection rate or decreasing the death rate by focusing on which target. Fundamentally, this incurs due to the inclination difference between age groups. Similarly, when there are specific groups that tend to have a higher immune system, their physical activities are likely to enlarge, leading to a higher infection rate for a total number of hosts increases. As a result, distributing vaccines through those factor analyses with computing the optimal ratio will serve as a key to balancing the casualties in the long run.

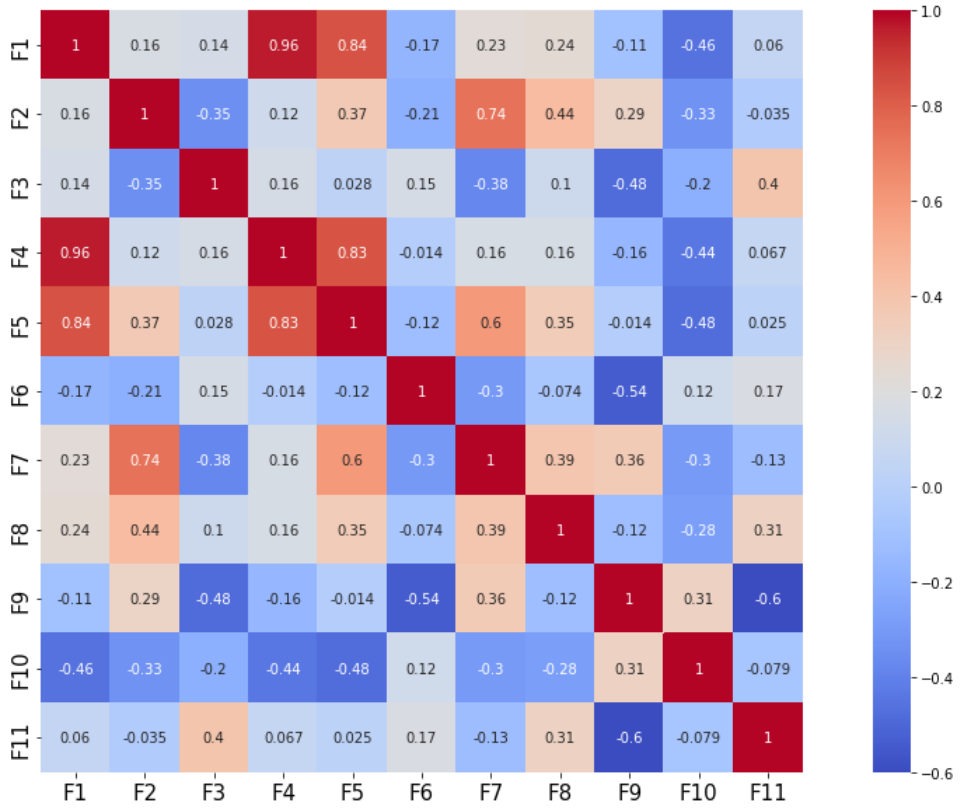


Figure 4.4 Pearson correlation coefficient of the combination of factors using Heatmap visualization

Based on the given features (F1~F11), our primary concern is to discover the significant relationship between the infection rate and the death rate, which may provide us with the feature-related statistical property while providing us the clue to comprehending the disease proliferation.

Figure 4.5 displays the density distribution, and $\sigma(infection\ rate) > \sigma(death\ rate)$ and several outliers exist that are higher than 3/2 times of upper quartile. This implies that the infection rate of the States is comparatively dispersed in a diverse range in general, which can be interpreted that it is highly being affected by diverse variables which differ by the properties of each State. In contrast, the death rate is denser with shorter distribution *x*-axis range, which asserts that it is not being strongly impacted than the infection rate. In addition, the average values in both factors show $\mu(infection\ rate) > \mu(death\ rate)$.

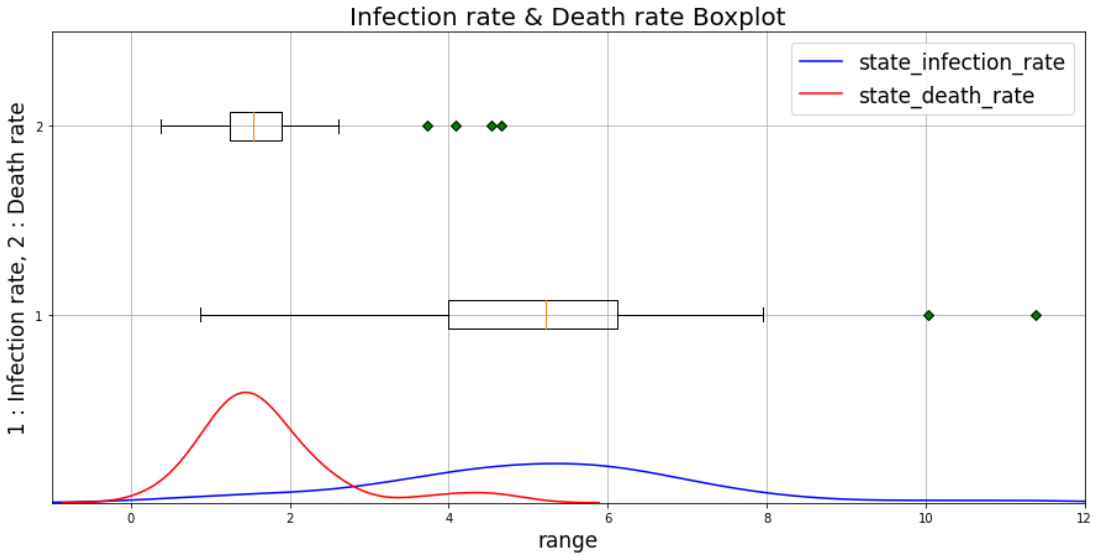


Figure 4.5 Distribution density of infection and death rate in 50 US states

Figure 4.6 shows the top five and bottom five states between the infection rate (*x*-axis) and other features (land area, GDP per capita, Median age, Population Density) of the *y*-axis. Figure 4.7 also displays the top five and bottom five states for the death rate (*x*-axis). From the *y*-axis, it has unsatisfying results without precise classification. Thus, we increase the function order higher than the first-order regression function.

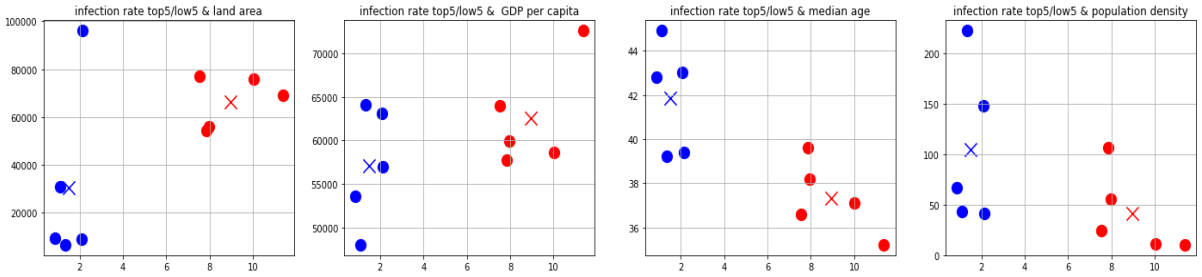


Figure 4.6. Top five and bottom five states in infection rates

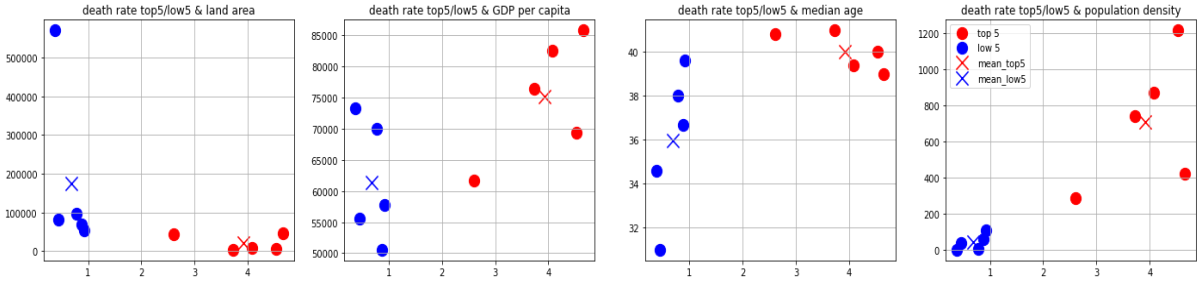


Figure 4.7 Top five and bottom five states in death rates

Figure 4.8 indicates the linear regression outcome (blue line), referring to the overall correlation between the two designated features. Figure 4.9 ~ 4.13 shows the polynomial regression based on the second-order, third-order, fourth-order, fifth-order, and tenth-order polynomial regression. Finding the order that evades the overfitting is critical, and since there are no further datasets that we can calibrate the *n*-th order of function, we compute their mean squared error and compare the result in Figure 4.14. In Figure 4.14, the coherency of the results is unstable, having a different phase of diminishing the error value. Generally, the error should exponentially decrease, with an urgent pace during the initial steps and slowing down at a certain point in the *x*-axis, which is the *n*-th order in the polynomial function. That specific point should be the optimal selection of order. For example, in MSE 7 in the subgraph of Figure 4.14, third-order is considered optimal. However, as the relationships show no coherent characteristic, features should be considered independently. It leads to the interpretation that considering features independently to reduce infection and death rates is practical.

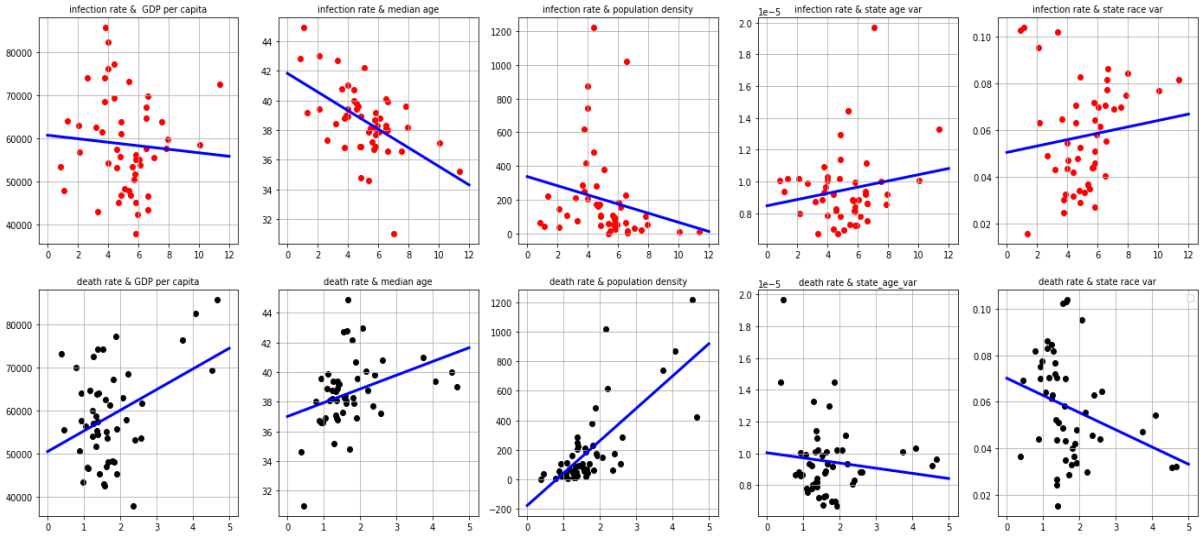


Figure 4.8 Linear regression visualization result of the infection rate (row 1) and death rate (row 2) dataset

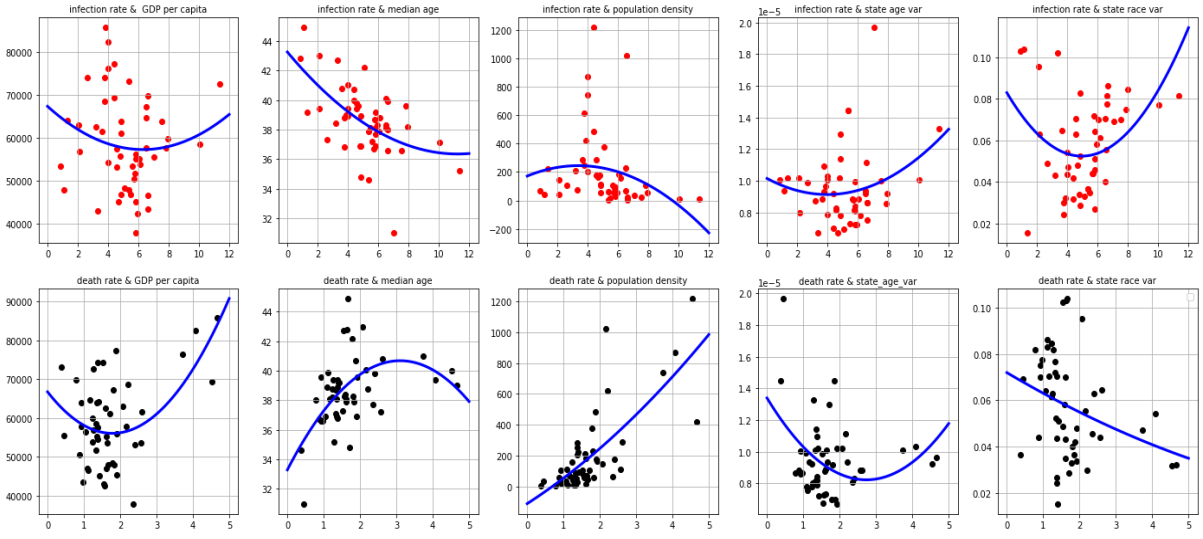


Figure 4.9 Second-order quadratic regression visualization result of the infection rate (row 1) and the death rate (row 2) dataset

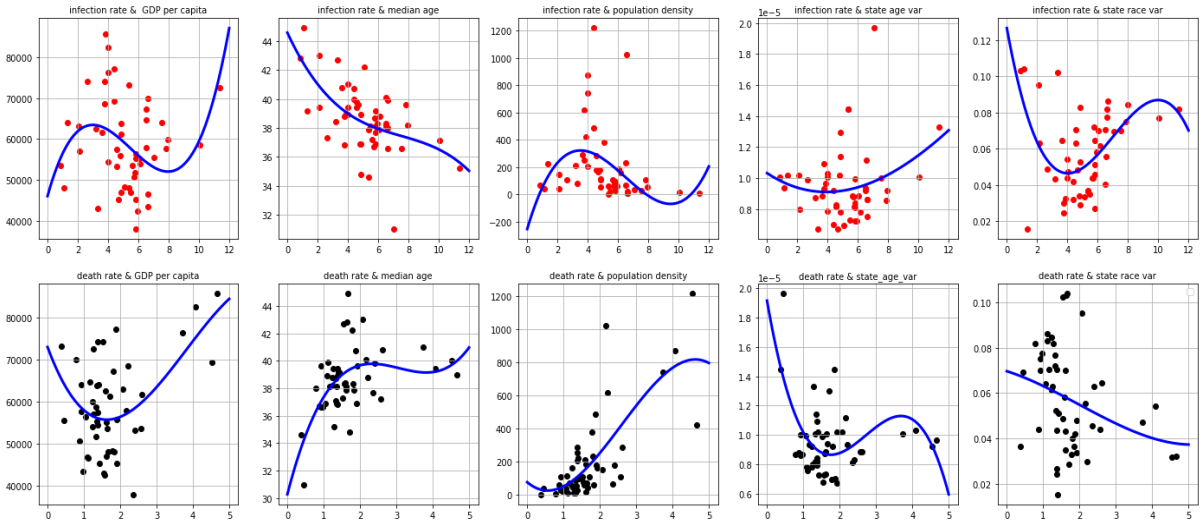


Figure 4.10 Third-order polynomial regression visualization result of the infection rate (row 1) and the death rate (row 2) dataset

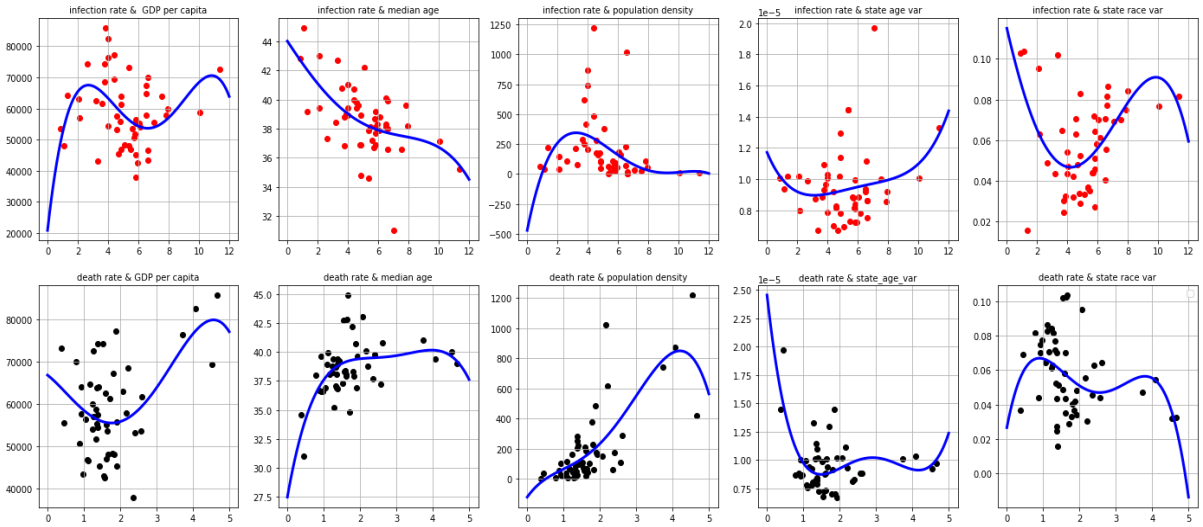


Figure 4.11 Fourth-order polynomial regression visualization result of the infection rate (row 1) and the death rate (row 2) dataset

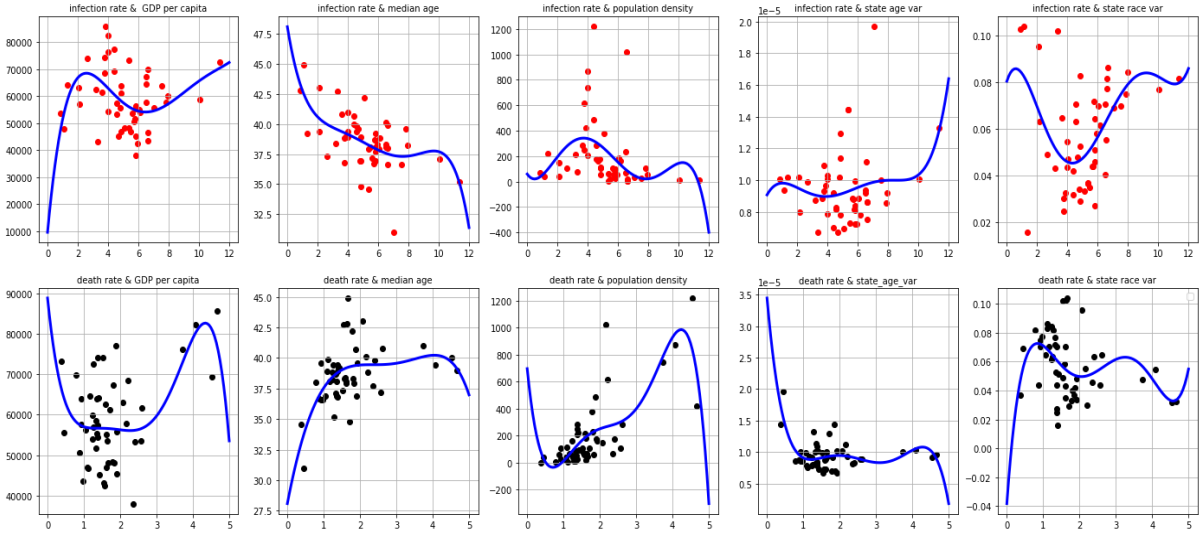


Figure 4.12 Fifth-order polynomial regression visualization result of the infection rate (row 1) and death rate (row 2) dataset

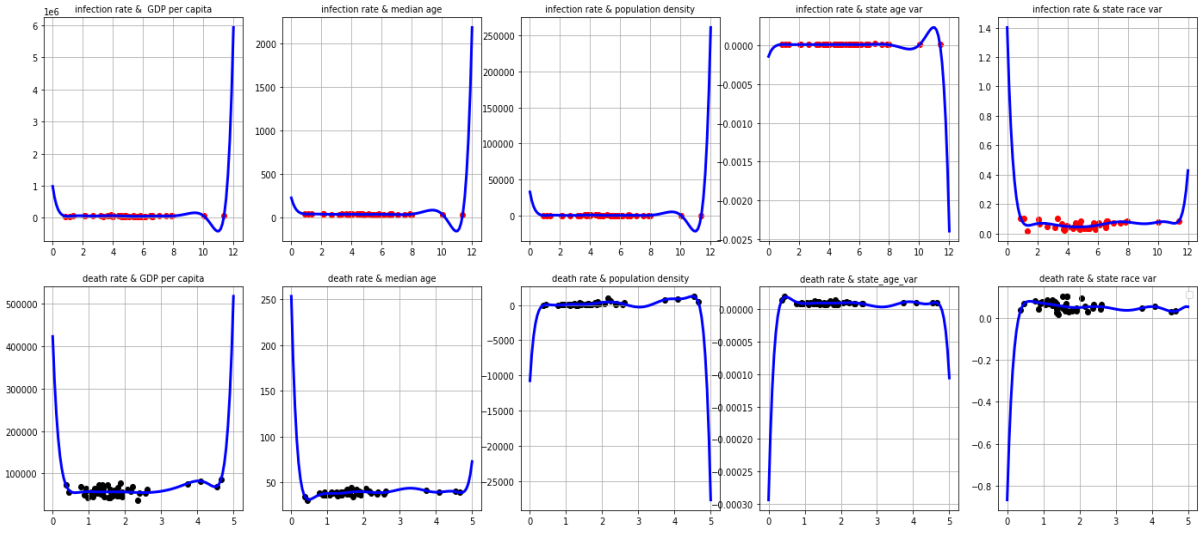


Figure 4.13 Tenth-order polynomial regression visualization result of the infection rate (row 1) and death rate (row 2) dataset

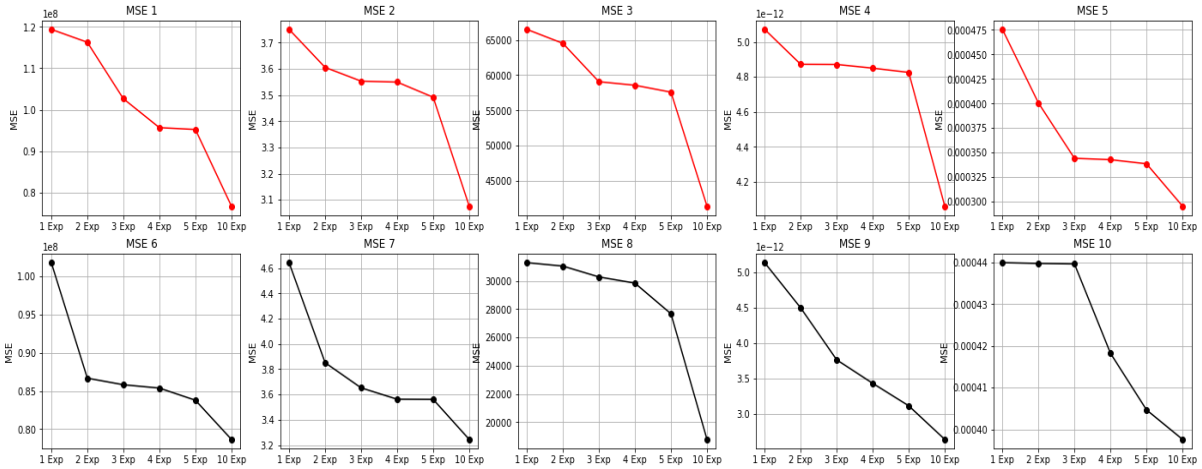


Figure 4.14 Mean Squared Error visualization in n-th polynomial regression result

5 EXPERIMENTS

In the experiment section, we conduct thorough experiments for validating the EVI by comparing it with the other criteria and indexes in the predetermined simulation environment and settings. Table 5.1 lists the assigned feature in each agent based on the proportion of statistics, where Mor refers to the mortality, and Iftr indicates the infection rate. Through the nine vaccination distribution trajectories, we evaluate three metrics: variation of infection cases, death cases, and the death rate. Based on the results, we reflect on the outcome in the real-practice and interpret the statistical phenomenon of the past COVID-19 cases.

Table 5.1 Calculated and assigned feature dataset for 300,000 nodes based on real-world statistics

Index	$\tilde{C}$	$\tilde{c}$	Age	Sex	R_1	R_2	R_3	…	Mor	ϕ	Iftr	EVI
1	$\tilde{C}_1$	$\tilde{c}_1$	0	F	$\mathbb{C}_{12}$	$\mathbb{C}_3$	$None$	…	0.164	0.238	0.712	0.261
2	$\tilde{C}_1$	$\tilde{c}_1$	1	M	$\mathbb{C}_7$	$None$	$\mathbb{C}_{19}$	…	0.238	0.408	0.364	0.402
…	…	…	…	…	…	…	…	…	…	…	…	…
300,000	$\tilde{C}_5$	$\tilde{c}_5$	85	M	$\mathbb{C}_{22}$	$\mathbb{C}_{21}$	$\mathbb{C}_5$	…	0.714	0.457	0.308	0.597

5.1 Experiment Settings

The vaccination simulation is mainly conducted through the highest to the lowest order of EVI, including eight other criteria (CVI [3], SVI [4], PVI [5], Age, Comorbidity risk, Age with Comorbidity risk, Random distribution, no vaccination) shown in Table 5.2. Note that each agent has its own unique corresponding index value. From the three metrics given, it is possible to infer the other indexes (*e.g.*, death cases + cured cases = infection cases); thus, we present those three metrics in the result. In the ABM, propagation methodology through EVI is explained in

pseudocode in Algorithm 1. Algorithm 1 functions similar to a stochastic random walk. Recall that this algorithm uniformly selects the node in binary method (*i.e.*, True or False) in each connected node from the source node. The probability of choosing, in other words, infecting the virus is uniform with a 50% probability. Recall that no precise data has been released for contracting the disease after the virus was injected. Therefore, we set the possibility of infection at 50% when there was a physical connection between the infected and the non-infected. The virus is dispersed to the connected sink node starting from the random source node, determining the infection based on the probabilistic value. The algorithm iteratively calls the recursion function in each time step, and the auxiliary buffer stores the status of each node (subject). If the node contracted the virus with a certain probability, it computes the death rate based on the assigned features and determines the status with a corresponding death rate.

Table 5.2 Vaccination simulation types

Simulation Type	
Type	Vaccine Distribution Type
Type 1	No Vaccination
Type 2	Random Vaccination
Type 3	Vaccination by Age
Type 4	Vaccination by Comorbidity Risk
Type 5	Vaccination by Age, Comorbidity Risk
Type 6	Vaccination by Social Vulnerability Index (SVI)
Type 7	Vaccination by COVID-19 Vulnerability Index (CVI)
Type 8	Vaccination by Pandemic Vulnerability Index (PVI)

Type 9 Vaccination by Epidemic Vulnerability Index (EVI)

Every agent has its status (Infected_Dead (*D*), Infected_Cured (*C*), Vaccinated (*V*), No virus (*A*)), and for efficient computation, we use matrix calculation that explains the status in the graph. We have equation (7) that assigns elements with $n_{(i,j)}$ having $n_{(i,j)} = EVI(n_{(i,j)})$ if $i = j$, and $n_{(i,j)} = e_{(i,j)}$ if $i \neq j$. This matrix was to compute the diagonal elements; $EVI(n_{(i,j)})$ and we define additional eight Matrixes ($M_2 \sim M_9$) that has the same size from M_1. For those eight Matrixes, the element of $n_{(i,j)}^{(o)} = n_{(i,j)}$ where $i \neq j$ and o denotes the matrix index. However, $n_{(i,j)}^{(o)}$ with $i = j$, which are diagonal elements they possess different values with No vaccination ($o = 2; M_2$), random vaccination ($o = 3; M_3$), vaccination by age ($o = 4; M_4$), vaccination by comorbidity risk ($o = 5; M_5$), vaccination by age + comorbidity risk ($o = 6; M_6$), vaccination by SVI [4] ($o = 7; M_7$), vaccination by CVI [3] ($o = 8; M_8$), vaccination by PVI [5] ($o = 9; M_9$).

We define another matrix M_s that has equivalent size ($300{,}000 \times 300{,}000$) and values of $n_{(i,j)}$ when $i \neq j$, but for $n_{(i=j,j=i)}$, we implement the status of each agent. For these diagonal elements, let $M(s)$ be $1 \times 300{,}000$ matrix size such that $M(s) = [\cup_{\forall i} status(n_{(i=j,j=i)})]$ in equation (9) where $(V, D) \in I$, c_I, c_D indicate the threshold on infection and death rate, $I(x)$ denotes the infection rate of input *x*, and $D(x)$ illustrates the death rate of input *x*. M_s offers to track the status of each agent and through which route it is being spread.

$$status(n_{(i=j,j=i)}) \begin{cases} 0 \ \ if\ status = A \\ 1 \ \ if\ status = V \\ tmp\ if\ status = I, tmp \begin{cases} -1\ if\ I(n_i) \geq c_I, & n_{(i=j,j=i)} \begin{cases} -1\ if\ D(n_i) \geq c_D \\ 1 \ \ if\ D(n_i) < c_D \end{cases} \\ 0 \ \ if\ I(n_i) < c_I \end{cases} \end{cases} \tag{9}$$

Each condition assigns the $status(n_{(i,j)})$, and ultimately, we have $M(s)$. With $M_{1\sim9}$, we sort the rows by the descending order of diagonal elements $n_{(i,j)}^{(o)}$. Based on sorting output, we start the simulation, releasing the vaccine distribution with the order of initial sequence with the specific amount per time unit.

To explain the notations used in Algorithm 1, R([a,b],w=c) indicates the random function that generates the random integer value between [a,b], with the biased selection based on the given hyperparameter(*i.e.*, weight) on choosing possible element. Also, $\ominus$ and $\oplus$ notation imply removing a designated element from the list and appending an element, respectively. For instance, $X \ominus x$ denotes that the element $x \in X$ is removed from the population list X, and $X \oplus x$ function indicates to add the element x at the end of list X. Simulations were performed with the 50% of infection rate when the edges touched the infected node to another, 20 initial spreaders were randomly given, and 500 vaccines were allocated per time unit. The simulations were demonstrated 100 times each in nine different criteria listed in Table 5.2. During each trial, the ABM was randomly reconstructed, generating with the formation of edge connection.

Algorithm 1: Vaccination through EVI

Input: Graph *g*, list of integers *patients*, list of strings *stat*, list of list *Dataset*, integer *vpt* (vaccine per time-unit)

Output: list of integers *time_unit*, list of strings *stat*

1 Initialization $(D, C, V) \leftarrow 0$, $(node\_lst, D\_lst, C\_lst, V\_lst) \leftarrow empty\ list$

2 **While** *V* ! = *vpt* **do**

3 | **if** *EVI_sort* = ∅ **then**

4 | | **Break**

```
    else
        if stat[EVI_sort[cnt]] = 'No Virus' then
            stat[EVI_sort[cnt]] ← 'V'     //vaccinated
            (V++) and (EVI_sort ⊖ EVI_sort[cnt])
for node ∈ {1,2, ... , n(patients)} do
    if stat[patients[node]] = 'No Virus' then
        if R([0,1], weight = [0.5, 0.5]) = 0 then   //infection rate 50%
            tmp ← R([0,1], weight = death rate in Dataset)
            if tmp = 0 then
                stat[patients[node]] ← 'D'      // Dead
                (D++) and (EVI_sort ⊖ EVI_sort[cnt])
            else
                stat[patients[node]] ← 'C'      // Cured
                (C++) and (EVI_sort ⊖ EVI_sort[cnt])
            for n ∈ {1,2, ... , # of neighbors of patients[node]} do
                node_lst ⊕ list of neighbors of patients[node][n]
(D_lst ⊕ D) and (C_lst ⊕ C) and (V_lst ⊕ V)
for node ∈ {1,2, ... , n(patients)} do
    for neighbor ∈ {1,2, ... , n(node_lst)} do
        if stat[node_lst[neighbor]] = 'D' or 'C' or 'V' then
            pass
        else
            time_unit ⊕ 'O'       // n(time_unit) = time passed
```

28			Vaccination through EVI(*g*, *patients*, *stat*, *Dataset*, *vpt*)

5.2 Experiment Result and Analysis

The cumulative result of death, cured, no virus, and vaccinated cases are shown in figure 5.1. In the figure, variations indicate the values are altered throughout the time and experience sharp conversions during the initial time periods. Figure 5.2 displays the acceleration; $|x_t - x_{(t+1)}|$ where *x* refers to the numerical value. The dynamics show the variance of the overall graph structure with respect to each criterion. The dispersion can mainly be divided into three phases: Increment, Decrement, and Stabilize. What we can deduce from this dispersion is the density of the graph. During the increment phase (10~30 time-steps), the infection and death cases surge until a certain point with $\frac{dC}{dt} > 0$ and $\frac{dD}{dt} > 0$ where C denotes the Cured and D indicates the dead. When $\lim_{t \to T} \frac{dD}{dt} = 0$, after $T + 1$, it shows $\frac{dC}{dt} < 0$ and $\frac{dD}{dt} < 0$. As it gradually becomes $\frac{dD}{dt} > \frac{dD}{dt'}$ where $t < t'$, we have $\lim_{t' \to T} \frac{dD}{dt'} = 0$ which becomes stabilized, with a linear variation in each factor. The variation graph entails a long-tail distribution format, which expands the right-side in Figure 5.2 (Note that Figure 5.2. is suggested to enlarge the view). The exponential increment of propagation asserts that the focus of the nodes is significant via edges, and in this time period, the contagious level is being maximized, following the equation (10), where $\forall < n_i, n_{i'} >$ indicates the for all neighbor edges that are connected to current node n_i.

$$\max Contagious\ Level = \max \bigcup_{\forall i} n(q_i) \quad s.t. \quad q_i = \{< n_i, n_{i'} > | i \neq i', (n_i, n_{i'}) \in g\} \tag{10}$$

The amount of infection cases during the incremental phase is $\sum_{t=10}^{30}(\text{number of nodes at } t) \approx 37{,}000$, which holds 12.33% for all populations. This value may not fit into the actual practice since we assumed that the subject becomes immune to the virus after single vaccination and also when naturally cured. In a real-world scenario, the immune system fades in time, and diverse conditions, as well as factors affect the infection, which leads to our problem into approximation.

To implement this into a real-world scenario, a single peak exists in each new COVID variant virus, whereas the period of the peak differs in each country based on diverse factors such as national disease control policy. For instance, the new variant Omicron (B.1.1.529), which was discovered in 2021 November, has a contagion rate of 613%, and around two months later, it surged the infection cases in the US. Currently, the statistics in the US have downsized, but other countries such as Japan and Korea are dealing with immense infection cases that they never experienced until now. Similarly, variant Delta+ (B.1.617.2.1) was found in 2021 June, and it struct the US during the following August, escalating the casualties. Therefore, each new variant causes an inevitable at least single impact surge due to its tolerance to a current vaccine and immune system or high infection rate. The overall trend is stabilized without any residual noise. This is due to the static graph structure, and since no viable benchmark dataset exists, expressing dynamic interactions had a minor limitation. Algorithm 1 runs on a graph model that is constructed as a real-world statistic. The numerical result throughout the time is analogous to the SEIR model [21], with indexes composing the entities with Susceptible, Exposed, Infected, and Recovered. In our simulation, recall that we compose the indexes with Cured (C), Dead (D) (Cured + Dead = Infected), No virus (A), and Vaccinated (V). With respect to the numerical result, the summation

would have 300,000 with the following equation (11), and the relationship of factors is shown in equation (12).

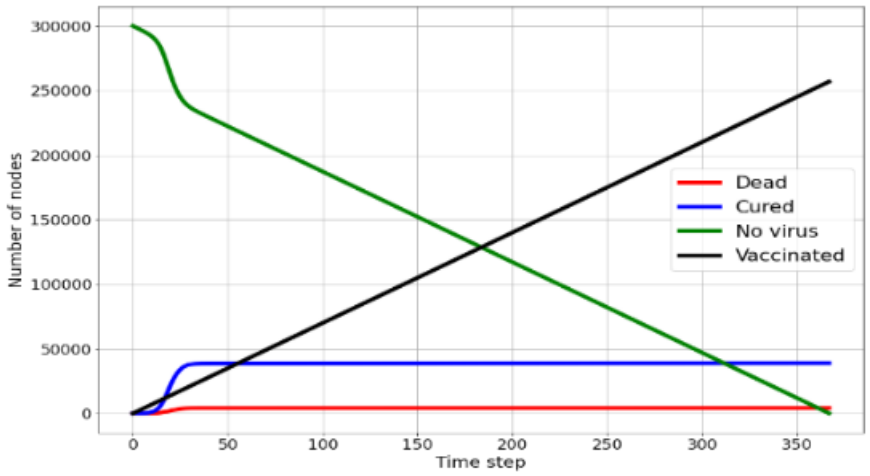


Figure 5.1 Cumulative visualization of four factors: Dead, Cured, No virus, and Vaccinated when distributing the vaccines through descending order of EVI

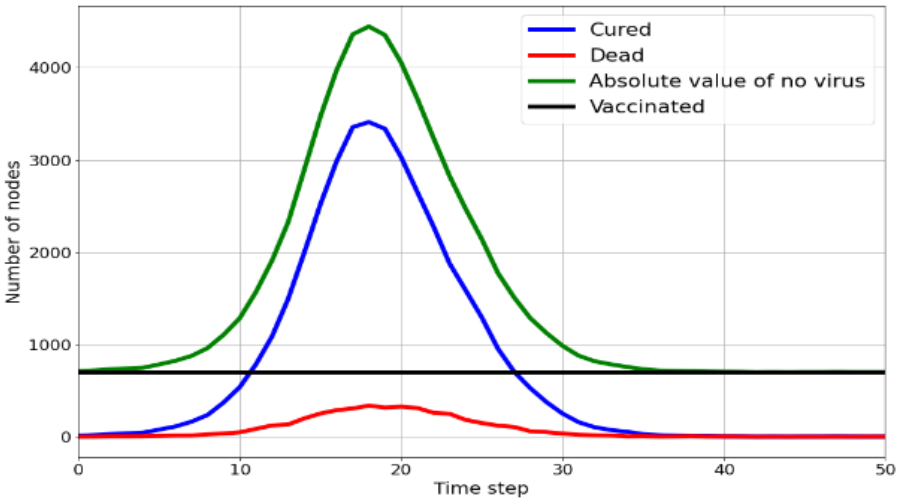


Figure 5.2 Variational visualization of four factors: Cured, Dead, Absolute value of no virus, and Vaccinated with time step 0~50 when distributing the vaccines through descending order of EVI

$$\sum_{t=0}^{T} C + \frac{dC}{dt} + \sum_{t=0}^{T} D + \frac{dD}{dt} + \sum_{t=0}^{T} V + \frac{dV}{dt} + \sum_{t=0}^{T} A + \frac{dA}{dt} = 300{,}000$$

(11)

$$\left|\frac{dC}{dt}\right| + \left|\frac{dD}{dt}\right| + \left|\frac{dV}{dt}\right| = \left|\frac{dA}{dt}\right|$$

(12)

The equation should be altered in an authentic world as (13) since subjects may be recontracted to the identical infection. In addition, the *dt* is larger than the single time unit since the maximum incubation period lasts more than a single time unit, for there exists a certain amount of period between the onset of the illness after the exposure. Furthermore, the advent of stronger variants increases the complexity, and analyzing whether variants are dependent or independent in parameter space or solution space takes much research.

$$\left|\frac{dC}{dt}\right| + \left|\frac{dD}{dt}\right| + \left|\frac{dV}{dt}\right| \approx \left|\frac{dA}{dt}\right|$$

(13)

Redesigned graph structure in every trial covers the diversity, and the following results are shown in Figures 5.3 ~ 5.8. The bar graph in Figures 5.3, 5.5, and 5.7 indicates the average value (bar) and the standard deviation (*i.e.*, black line in the center) of each criterion: death cases, infection cases, and death rate after the 100 trials of simulations. In every figure, 'no virus' criteria were omitted since its infection reached almost 100% (=1.0), and the decimal points in every figure illustrate the percentage of the infection compared to the no virus result. For example, the EVI value in a death case has 0.403, which is equivalent to 40.3%, compared to the no vaccination case has 100%. Additionally, the quartiles of the density graph of all the trials are displayed in Figures 5.4, 5.6, and 5.8, where the green rhombus shape value indicates the statistical outlier. In most cases, EVI incorporates the lower casualties: an average of 9.4% lower in death cases, 5.0% lower in infection cases, and 3.5% lower in death rates than other distribution criteria.

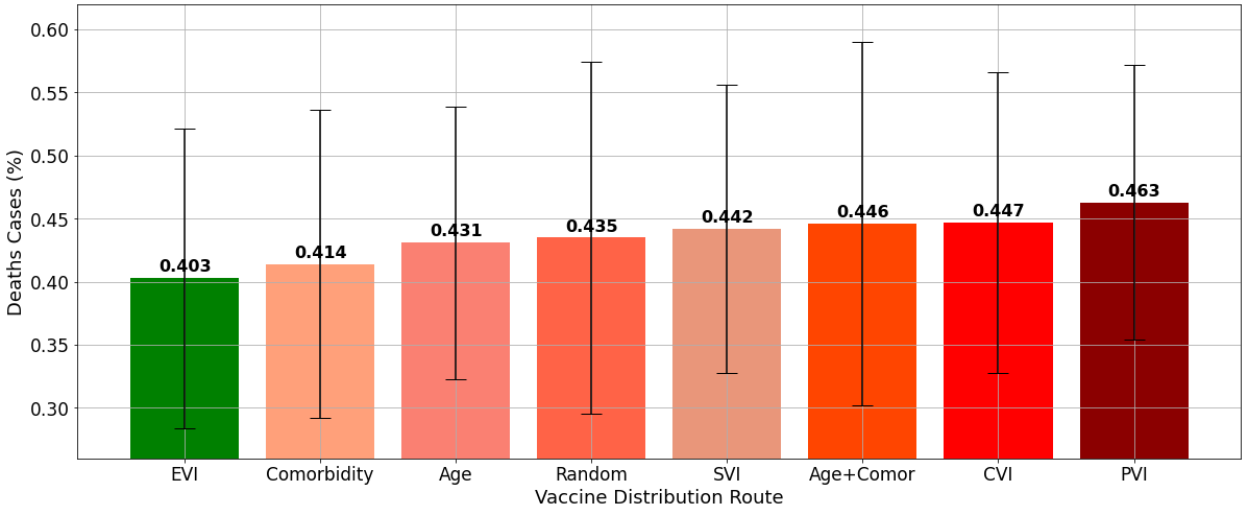


Figure 5.3 Death case results when distributing the vaccines based on each criteria

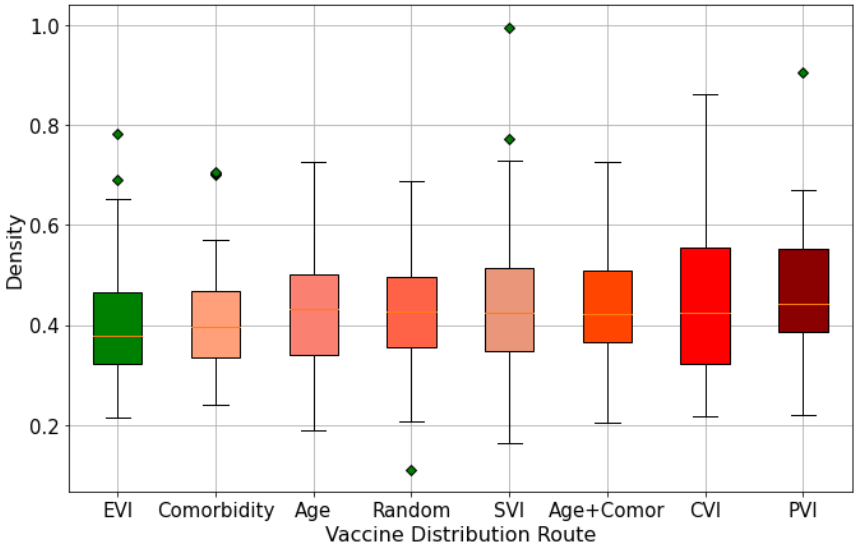


Figure 5.4 Boxplot showing the distribution after 100 simulation trials in death cases

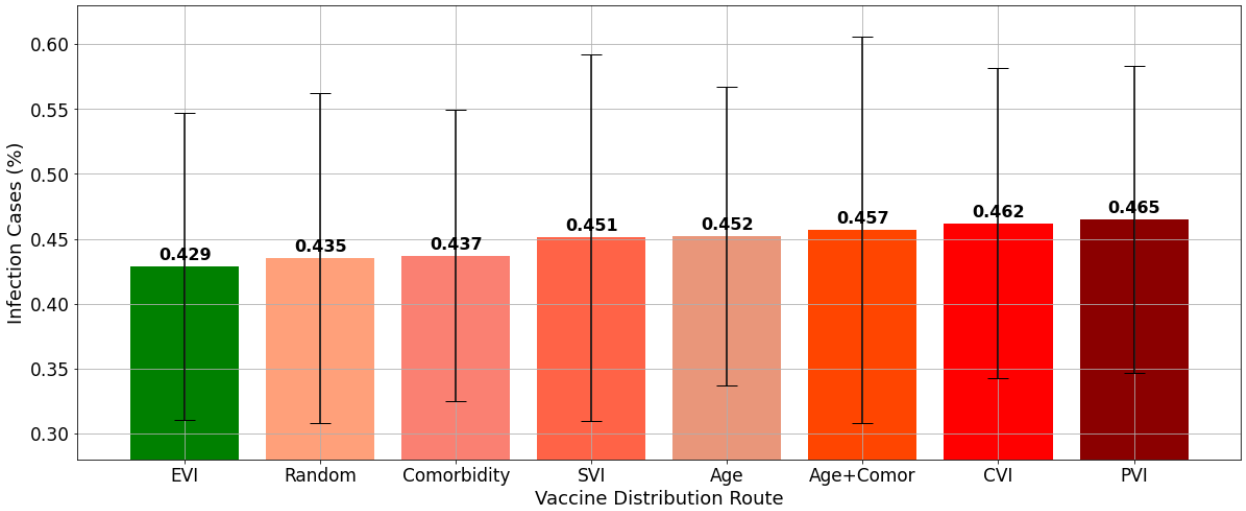


Figure 5.5 Infection case results when distributing the vaccines based on each criteria

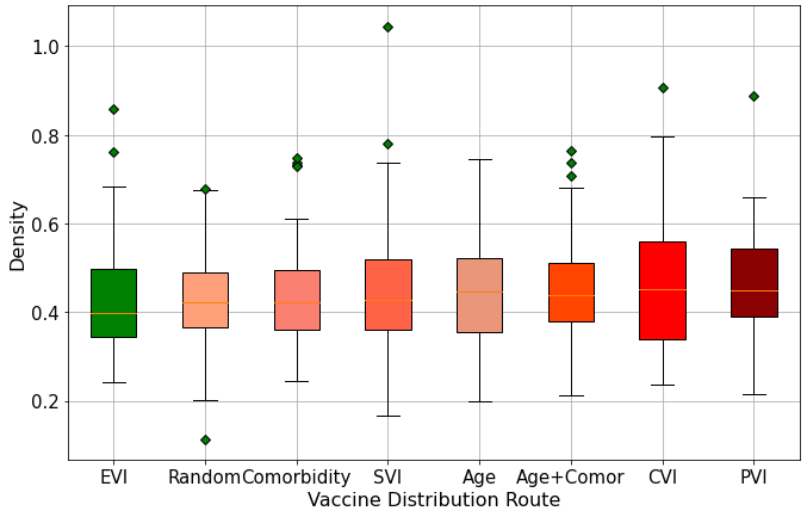


Figure 5.6 Boxplot showing the distribution after 100 simulation trials in infection cases

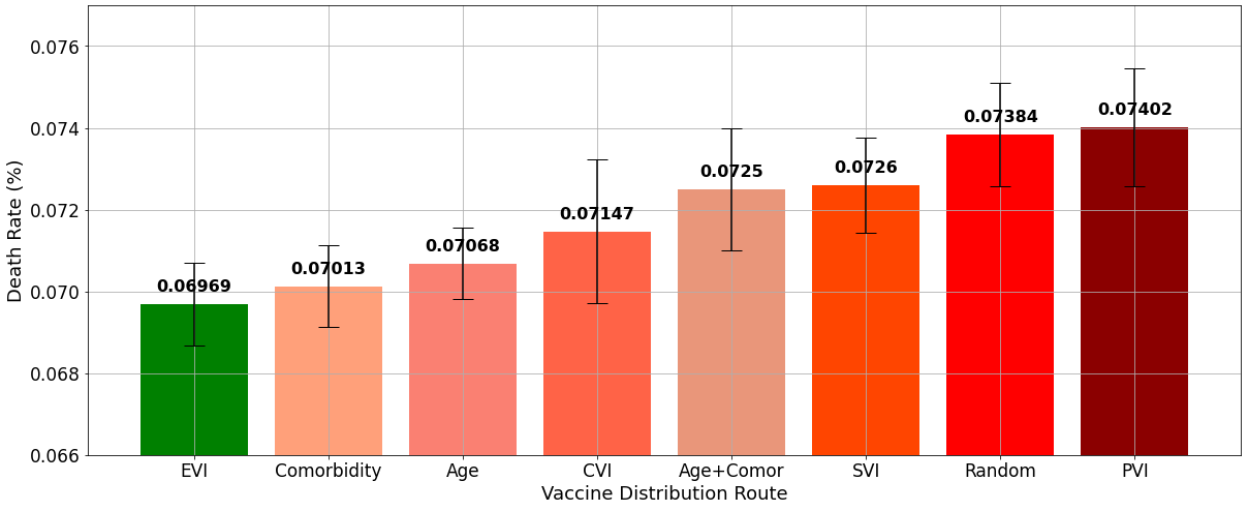


Figure 5.7 Death rate results when distributing the vaccines based on each criteria

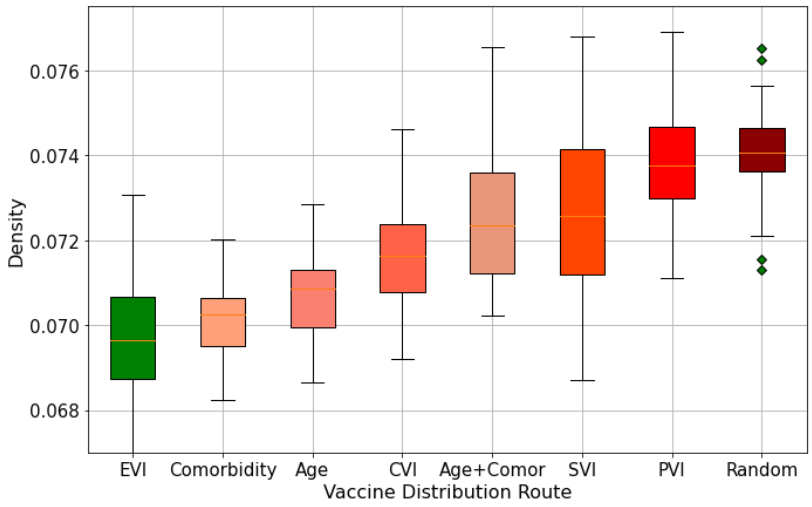


Figure 5.8 Boxplot showing the distribution after 100 simulation trials in death rate

6 CONCLUSION

This thesis paper proposes the novel Epidemic Vulnerability Index, which is an optimal criterion that estimates the potential threat of the uninfected subject, utilizing the internal clinical attributes and external social attributes. Determining the routes of the vaccine propagation is a sensitive task that requires thorough analyses concerning the impact with respect to various conditions such as age, region, infection rate, death rate, etc. Through our analysis, we have estimated the potential risk that the subject incorporates based on their biological and social factors and shown that a trade-off exists when distributing vaccines, for there is a negative correlation between the infection rate and the death rate. This is a severe dilemma that attempts to diminish the death rate; when concentrating the vaccination on the elderly person also, subjects with high comorbidities would indeed decrease the death rates but increment the infection cases. On the contrary, controlling the infection rate that targets the vaccination to the younger generation would enhance the current death rate. Our solution to this problem is to adaptively control the weights of both cases based on the present and predicted trends via EVI. In order to validate the performance of the EVI, this thesis proposes a simulation-based inference that estimates the impact of the vaccination scenario. Through the graph-structured Agent-Based Model (ABM) that was formalized based on real-world statistics, we assign the features to every 300,000 agents representing the community. In the ABM, we conduct multiple virus & vaccine proliferation simulations with nine scenarios, including the vaccination through the descending order of EVI. The simulation is similar to the Markov Chain Monte Carlo simulation, where it stochastically disperses the virus and injects the vaccines to the optimal agents and observes the variations of the metrics of infection cases, death cases, cured cases, vaccinated cases, and no-virus cases. Compared to the existing indexes and other vaccination routes, vaccination via EVI has shown to

have a 5.0% lower number of infection cases, 9.6% lower number of death cases, and 3.5% lower death rates.

EVI is designed to integrate the different propensity of our two main targets that must be controlled: infection rate and death rate. In addition, the evaluation has shown to be effective in the ABM through empirical simulations of various vaccination scenarios. However, dynamics in epidemiology require complex modeling. The environmental system that we currently live in consists of multivariate factors and complicated issues that cannot be predicted, including randomness, dynamic and heterogeneous entities, vaccine performance, etc. Estimating the future variations with high accuracy is an arduous task, and our limitation dwells in this domain. Especially, the ABM in this thesis has static factors compared to the real-practice system that evolves dynamically. Therefore, constructing the ABM analogous to the real-world system would overcome the limitations and enhance the accuracy.

In reality, the decision-making process analyzes the feasible candidates and selects the options in accordance with the response to the current and future situation. It combines multiple strategies, establishing a method that covers the various problem. Likewise, this study can be applied when constructing the successful vaccination for other future pandemics that may occur. It could also be utilized when predicting the medical supply in the region, such as regional vaccine distribution scenarios, via adopting and finetuning novel attributes. Ultimately, these endeavors would come as obliging studies when organizing the optimal response strategy for the pandemic at the right time and place.

REFERENCE

[1] World Health Organization, Accessed: Mar. 15, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

[2] World Health Organization. “WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020”, Accessed: Mar. 15, 2020. [Online]. Available: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

[3] D. Decaprio et al., “Building a COVID-19 vulnerability index,” Journal of Medical Artificial Intelligence, vol. 3, 2020.

[4] B. E. Flanagan, E. W. Gregory, E. J. Hallisey, J. L. Heitgerd, and B. Lewis, “A social vulnerability index for disaster management,” Journal of homeland security and emergency management, vol. 8, no. 1, 2011.

[5] S. W. Marvel et al., “The COVID-19 Pandemic Vulnerability Index (PVI) Dashboard: Monitoring county-level vulnerability using visualization, statistical modeling, and machine learning,” Environmental Health Perspectives, vol. 129, no. 1, 2021.

[6] O. Amram, S. Amiri, R. B. Lutz, B. Rajan, and P. Monsivais, “Development of a vulnerability index for diagnosis with the novel coronavirus, COVID-19, in Washington State, USA,” Health & place, vol. 64, no. 4, 2020.

[7] A. A. Salisu, and O. A. Lateef, “Constructing a global fear index for the COVID-19 pandemic,” Emerging Markets Finance and Trade, vol. 56, no. 10, pp. 2310—2331, 2020.

[8] H. Pinglin et al., “Accounting index of COVID-19 impact on Chinese industries: A case study using big data portrait analysis,” Emerging Markets Finance and Trade, vol. 56, no .10, pp. 2332—2349, 2020.

[9] S. D. Searle, A. Mitnitski, E. A. Gahbauer, T. M. Gill, and K. Rockwood, “A standard procedure for creating a frailty index,” BMC geriatrics vol. 8, no.1, pp. 1—10, 2008.

[10] G. Bellelli, R. Paola, and G. Citerio, “The role of frailty in COVID-19 patients,” Intensive Care Medicine, vol. 46, no. 10, pp. 1958—1959, 2020.

[11] G. Bellelli et al., “Frailty index predicts poor outcome in COVID-19 patients,” Intensive care medicine, vol. 46, pp.1634—1636, 2020.

[12] K. M. Bubar et al., “Model-informed COVID-19 vaccine prioritization strategies by age and serostatus,” Science, vol. 371, no. 6532, pp. 916-921, 2021.

[13] L. Corey, J. R. Mascola, A. S. Fauci, and F. S. Collins, “A strategic approach to COVID-19 vaccine R&D,” Science, vol. 368, no. 6494, pp. 948—950, 2020.

[14] A. R. Tuite, L. Zhu, D. N. Fisman, and J. A. Salomon, “Alternative dose allocation strategies to increase benefits from constrained COVID-19 vaccine supply,” Annals of internal medicine, vol. 174, no. 4, pp. 570—572, 2021.

[15] C. C. Kerr et al., “Covasim: an agent-based model of COVID-19 dynamics and interventions,” PLOS Computational Biology, vol. 17, no. 7, 2021.

[16] Centers for Disease Control and Prevention, 2009 H1N1 Flu (2009); Accessed: Nov. 11, 2021 [Online], Available: www.cdc.gov/h1n1flu/vaccination/vaccinesupply.html.

[17] A. Coustasse, C. Kimble, and K. Maxik, “COVID-19 and Vaccine Hesitancy: A Challenge the United States Must Overcome,” The Journal of Ambulatory Care Management, vol. 44, no. 1, pp. 71—75, 2021.

[18] J. Wang, Y. Peng, H. Xu, Z. Cui, and R. O. Williams, “The COVID-19 vaccine race: challenges and opportunities in vaccine formulation,” AAPS PharmSciTech, vol. 21, no. 6, pp. 1—12, 2020.

[19] Y. A. Adebisi, G. I. Oke, P. S. Ademola, I. G. Chinemelum, I. O. Ogunkola, and D. E. Lucero-Prisno. “SARS-CoV-2 diagnostic testing in Africa: needs and challenges,” The Pan African Medical Journal, vol. 35, no. 4, 2020.

[20] M. Rastegar, M. Tavana, A. Meraj, and H. Mina, “An inventory-location optimization model for equitable influenza vaccine distribution in developing countries during the COVID-19 pandemic,” Vaccine, vol. 39, no. 3, pp. 495—504, 2020.

[21] S. Mwalili, M. Kimathi, V. Ojiambo, D. Gathungu, and R. Mbogo, “SEIR model for COVID-19 dynamics incorporating the environment and social distancing,” BMC Research Notes, vol. 13, no. 1, pp. 1—5, 2020.

[22] L. Matrajt, J. Eaton, T. Leung, and E. R. Brown, “Vaccine optimization for COVID-19: Who to vaccinate first?,” Science Advances, vol. 7, no. 6, 2021.

[23] C. Wolfram, “An agent-based model of COVID-19,” Com-plex Systems, vol. 29, no. 1, pp. 87—105, 2020.

[24] M. S. Shamil, F. Farheen, N. Ibtehaz, I. M. Khan, and M. S. Rahman, “An agent-based modeling of COVID-19: validation, analysis, and recommendations,” Cognitive Computation, pp. 1—12, 2021.

[25] P. CL. Silva et al., “COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions,” Chaos, Solitons & Fractals, vol. 139, 2020.

[26] J. Gomez, J. Prieto, E. Leon, and A. Rodriguez, “INFEKTA: a general agent-based model for transmission of infectious diseases: studying the COVID-19 propagation in Bogotá-Colombia,” PloS one, vol. 16, no. 2, 2021.

[27] N. Hoertel et al. “Facing the COVID-19 epidemic in NYC: a stochastic agent-based model of various intervention strategies,” MedRxiv (2020).

[28] R. Hinch et al., “OpenABM-Covid19—An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing,” PLoS computational biology, vol. 17, no. 7, 2021.

[29] S. Y. Del Valle, J. M. Hyman, H. W. Hethcote, and S. G. Eu-bank, “Mixing patterns between age groups in social networks,” Social Networks, vol. 29, no. 4, pp. 539—554, 2007.

[30] M. M. Hughes, “County-Level COVID-19 Vaccination Coverage and Social Vulnerability,” Morbidity and mortality weekly report, vol. 70, no. 12, pp. 431—436, 2021.

[31] K. M. Gostic, et al., “Practical considerations for measuring the effective reproductive number”, R t. PLoS computational biology, vol. 16, no. 12, 2020.

[32] U. Nguemdjo, F. Meno, A. Dongfack, and B. Ventelou, “Simulating the progression of the COVID-19 disease in Cameroon using SIR models”, PLoS One, vol. 15, no. 8, 2020.

[33] A. Ajbar, R. T. Alqahtani, and M. Boumaza, “Dynamics of an SIR-based COVID-19 model with linear incidence rate, nonlinear removal rate, and public awareness”, Frontiers in Physics, pp. 13, 2021.

[34] R. T. Alqahtani, “Mathematical model of SIR epidemic system (COVID-19) with fractional derivative: stability and numerical analysis”, Advances in Difference Equations, vol. 1, pp. 1-16, 2021.

[35] S. A. Alanazi, M. M. Kamruzzaman, M. Alruwaili, N. Alshammari, S. A. Alqahtani, and A. Karime, “Measuring and preventing COVID-19 using the SIR model and machine learning in smart health care”, Journal of healthcare engineering, 2020.

[36] Y. C. Chen, P.E. Lu, C. S. Chang, and T. H. Liu, “A time-dependent SIR model for COVID-19 with undetectable infected persons”, IEEE Transactions on Network Science and Engineering, vol. 7, no. 4, pp. 3279-3294, 2020.

[37] J. H. Son, H. J. Hwang, and S. Y. Jung, “Analysis of COVID-19 spread in South Korea using the SIR model with time-dependent parameters and deep learning”, medRxiv, 2020.

[38] V. Ram, and L. P. Schaposnik, “A modified age-structured SIR model for COVID-19 type viruses. Scientific Reports”, vol. 11, no. 1, pp. 1-15, 2021.

[39] S. Moein, et al., “Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan”, Scientific Reports, vol. 11, no. 1, pp. 1-9, 2021.

[40] N. D. Yanez, N. S. Weiss, J. A. Romand, and M. M. Treggiari, “COVID-19 mortality risk for older men and women,” BMC Public Health, vol. 20, no.1, pp. 1—7, 2020.

[41] S. Mukherjee, and K. Pahan, “Is COVID-19 Gender-sensitive?,” Journal of Neuroimmune Pharmacology, vol. 16, no. 4, pp. 38—47, 2021.

[42] Centers for Disease Control and Prevention, Accessed: Apr. 12, 2021. [Online], Available: https://data.cdc.gov/widgets/hk9yquqm.

[43] NYTimes, Accessed: Apr. 06, 2021. [Online], Available: https://github.com/nytimes/covid-19-data.

[44] Medscape, “Almost 90% of COVID-19 Admissions Involve Comorbidities,” Accessed: Apr. 18, 2021. [Online], Available: https://www.medscape.com/viewarticle/928531.

[45] A. Sanyaolu et al., “Comorbidity and its impact on patients with COVID-19,” SN comprehensive clinical medicine, pp. 1—8, 2020.

[46] B. Wang, R. Li, Z. Lu, and Y. Huang, “Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis,” Aging (Albany NY), vol. 12, no. 7, pp. 6049—6057, 2020.

[47] S. Jakhmola et al., “Comorbidity assessment is essential during COVID-19 treatment,” Frontiers in physiology, vol. 11, 2020.

[48] F. Xiaoyu et al., “Epidemiological, comorbidity factors with severity and prognosis of COVID-19: a systematic review and meta-analysis,” Aging (Albany NY), vol. 12, no. 13, pp. 12493—12503, 2020.

[49] World Health Organization, Accessed: Apr. 08, 2021. [Online], Available: https://www.who.int/standards/classifications/classification-of-diseases.

[50] A. H. Dekker, “Network centrality and super-spreaders in infectious disease epidemiology,” in 20th Int. Congress on Modelling and Simulation (MODSIM2013). Adelaide, South Australia, 2013.

[51] G. Sabidussi, “The centrality index of a graph,” Psychometrika, vol. 31, no. 4, pp. 581—603, 1966.

[52] D. Sharma, A. Surolia, W. Dubitzky, O. Wolkenhauer, K. H. Cho, and H. Yokota, “Degree Centrality,” eds. New York: Encyclopedia of Systems Biology. Springer, 2013.

[53] L. Leydesdorff, “Betweenness centrality as an indicator of the interdisciplinarity of scientific journals,” Journal of the American Society for Information Science and Technology, vol. 58, no. 9, pp. 1303—1319, 2007.

[54] P. D. Straffin, “Linear algebra in geography: eigenvectors of networks,” Mathematics Magazine, vol. 53, no. 5, pp. 269—276, 1980.

[55] D. Sullivan, “What is Google PageRank? A guide for searchers & webmasters,” Search engine land, 2007.

[56] Statistca, Accessed: Aug. 15, 2021. [Online], Available: https://www.statista.com/statistics/241488/population-of-the-us-by-sex-and-age/

[57] D. Weycker et al., “Population-wide benefits of routine vaccination of children against influenza,” Vaccine, vol. 23, no.10, pp. 1284–-1293, 2005.

[58] J. Medlock, and A. P. Galvani, “Optimizing influenza vaccine distribution,” Science, vol. 325, no.5948, pp. 1705—1708, 2009.

[59] S. Bansal, B. Pourbohloul, and L. A. Meyers, “A comparative analysis of influenza vaccination programs,” PLOS Med. Vol. 3, 2006.

[60] H. Lee, et al., “Epidemic Vulnerability Index for Effective Vaccine Distribution Against Pandemic,” International Symposium on Bioinformatics Research and Applications. Springer, Cham, pp. 22—34, 2021.