Agent Based Models for Optimizing Global COVID-19 Vaccine Allocation

Qingfeng Li[1], Yajing Huang[1]

[1]Department of International Health, Johns Hopkins University Bloomberg School of Public Health, Baltimore MD 21205, USA

Correspondence should be addressed to: qli28@jhu.edu

Abstract

xx

1 Introduction

Vaccination is key to ending the global COVID pandemic that has infected xxx million people and caused x million deaths globally as of mm/dd/yyyy [1]. However, there will not be enough COVID vaccines for all people in every country in the near future (citation). As a result, there is a need to design vaccine allocation strategies among countries and across population groups within countries.

Allocating vaccines across populations and countries is a complex task and has to consider many demographic and pandemic factors. Optimal allocations need to be identified through either real-world or virtual experiments.

In this article, we present agent-based models, called Global-Vax, that can be used to evaluate and compare current vaccine allocation strategies and explore alternative approaches. Section 2 describes its modeling framework and components, including individual agent properties and population characteristics, epidemiological progression model, regional and international transmission dynamics and vaccination implementation. Section 3 illustrates how Global-Vax is parameterized and validated using COVID-19 data, then it is projected into the future to identify optimal vaccination allocation strategies in terms of equity and effectiveness. Section 4 discusses the limitations of Global-Vax and future improvements.

2 Modeling

2.1 Overview

The Global-Vax model is designed to simulate the dynamics of the COVID pandemic at global, regional, and individual levels (Figure 1). It is a meta-population model based on a collection of interconnected regional ABMs. On the global level, viral transmission occurs via international travels. Connections between regional models are made via importations of infections cases. The regional models are either at the national level, or at a provincial/state level, within which agents interact and make contacts with one another. An agent represents a person with an epidemiological status and demographic attributes. The state transition of an agent is governed by an extension to the Susceptible-Infectious-Removed (SIR) model. The possible states include susceptible (S), asymtomatically infected (A), infected with symptoms (I),recovered (R), and dead (D).

The Global-Vax model evolves via an iterative process. As described in Algorithm 1, at the very beginning, we initialize all three levels of the model: individual agents a, regional ABM m, and the meta-population model $\mathcal{M}$. We also use a dataframe Θ to keep track of the regional numbers of agents in each epidemiological status. At each time step (one day), the procedure is as follows. At first, based on the COVID prevalence of neighboring regions as of the last time step, the number of cases infected by imported cases is computed and they are added to each regional model (Section 2.3). Within each regional model, agents interact and viral transmission occurs (Section 2.5). Next, some agents are randomly chosen to be vaccinated, based on a given regional vaccination plan (Section 2.6). Then the state of each agent is updated based on their demographic and statistical properties, as well as their vaccination statues (Section 2.4, 2.6). At last, the model records the total number of agents in a specific state and the number of confirmed new cases in a region.

Algorithm 1 Procedure of the Global-Vax model

Require: P the parameter set, T the number of steps

```
a, m, M, Θ ← initialize(P)                          ▷ a ∈ m ∈ M
for t ← 1 to T do
    for all model m_r ∈ M do
        m_r.import_infectious_cases(Θ, P)
        for all agent a_i ∈ m_r do
            if a_i is infectious then
                a_i.initiate_meetings(P)
            end if
        end for
        m_r.vaccinate(P)
        for all agent a_i ∈ m_r do
            a_i.update_state(P)
        end for
    end for
    Θ.collect_data()
end for
```

In reality, we do not (or cannot) usually observe all the numbers of agents that are in the five compartments S, A, I, R, D. The numbers that often get reported are daily new cases, cumulative cases, daily new deaths and cumulative deaths. We assume that only symptomatic infections can be observed in real life. Thus in our simulations, we keep track of the number of agents newly entering compartment I at each step as the daily new cases I_{new}, and the sum over all new cases in the past as cumulative cases C_I. I_{new} and C_I are stored in the dataframe Θ.

The code is structured such that the regional ABM is implemented by an object called **ABM**, with a collection of agents whose information is stored in Numpy arrays, for fast lookup and filtering. The **ABM** object is responsible for the majority of tasks including disease transmission and progression, as well as vaccination. It can run independently as an isolated region. This allows for parallelization during parameter fitting (Section 3), reducing time complexity significantly. The meta-population model is implemented by a **Meta** object, which consists a list of **ABM**s. It mainly handles interconnections between **ABM** objects.

The **ABM** object can be used to simulate an epidemic in one country or one province/state. In reality, the population sizes of different regions range approximately from tens of thousands to billions. Models with all other parameters held fixed but vary in population sizes can behaved very differently. Moreover, modeling a very large population is computationally intensive. Therefore, we fix the population N_{pop} of all regional **ABM**s to be one million, with an additional parameter S_{pop} defined as the ratio of the region's actual populations to one million. The output from **ABM**, e.g. daily new cases, will be scaled in the end with S_{pop}.

Important parameters for simulations are listed in Table 1.

2.2 Definitions

2.2.1 Agents

The regional ABM is made up of dynamic agents $a_i(t)$, for $i = 1,\ 2, ...\ N$, where N is the total number of agents in the model. $a_i(t)$ denotes its epidemiological state at time t. $a_i(t)$ is categorical, with categories 1, 2, 3, 4, 5 referring to state S, A, I, R, D respectively. We also use T_{s_1} to the days an agent stays in state s_1 or $T_{s_1 \to s_2}$ if the next state s_2 can be specified. T_{s_1} and $T_{s_1 \to s_2}$ are to be drawn from probability distributions.

For each agent i, x_i denotes its demographic attributes, including sex (male, female), age (children, adults, and seniors), and race (majority, minority). The probabilities of infection via contacts and disease fatality depend on x_i. At the moment, the number of demographic groups is 2 sex $\times$ 3 age $\times$ 2 race = 12. That means x_i is a categorical variable ranging from 1 to 12.

Under certain rules, the agents can perform the following behaviors: meet with other agents or obtain vaccination .

We use π_i to denote an agent's preference to meet other agents. With $M_{\max}$, the maximum number of meetings initiated by any agent, $\pi_i \sim \text{Uniform}(0, M_{\max})$ (discreet), is the number of agents agent i meets in a day.

In addition, we use $\mathcal{I}_i$ to denote the infectiousness of agent i. Agents at state I are always infectious. For agents in state A, they become infectious 3 days after transitioning to A, if

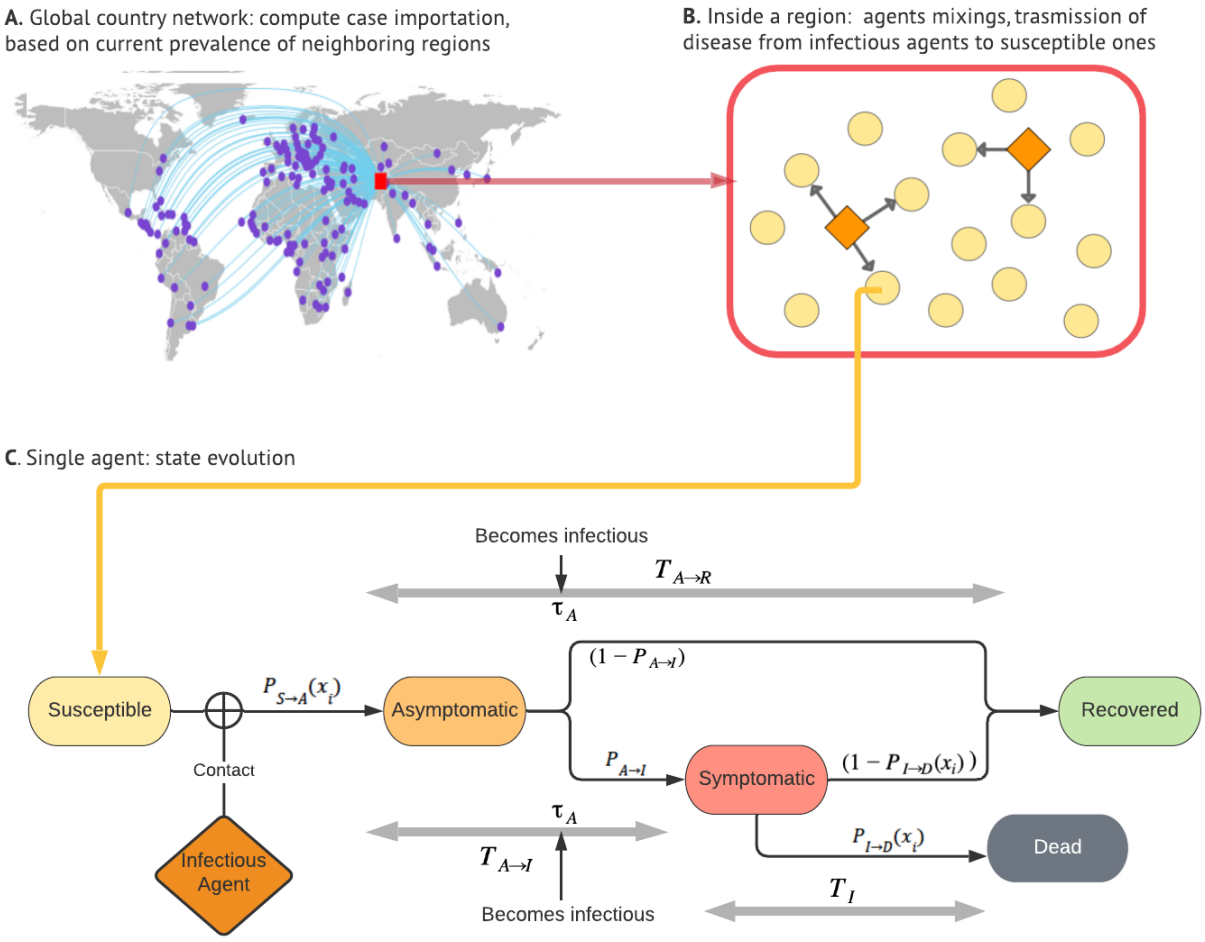

Figure 1: Simulation Framework

Table 1: Simulation model parameter values and sources

Epidemiological			
Parameter	Description	Values or distribution	Reference
$P_{S\to A}$	Transmission probability per contact of senior females	[0.006,0.02]	Estimated from parameter fitting
$S(x_i)$	Scale factor for $P_{S\to A}$, transmission probability relative to that of senior females (which means transmission probability per demo-group is)	[0,1]	Estimated from COVID sex and age disaggregated data of [20]
$P_{A\to I}$	Proportion of infected agents that eventually show symptoms	0.45	[21]
$T_{A\to I}$ $T_{A\to R}$	Duration of asymptomatic period until symptoms start to show (if the next state is I), or until agents recover (if the next state is R)	If the next state is I, lognormal($1.63,0.5^2$); If the next state is R, lognormal($2.23,0.2^2$); Cut off at 21 days.	[22], [23]
τ_A	The day when an asymptomatic case become infectious	If the next state is I, 2 days before turning into I, or day 1; If the next state is R, day 3.	[24], [25], [23]
$P_{I\to D}(x_i)$	Case fatality of symptomatic cases	[0,0.1]	[20]
T_I	Duration of symptomatic infection	lognormal($2.04,0.4^2$); Cut off at 21 days.	[26]
Additional parameters for the interaction of agents			
π_i	The number of contacts an infectious agent has daily	Uniform(0,M_{max})	M_{max} is assumed to be 30
λ_M	The scaling down factor for the number of contacts daily of agents that are infected and symptomatic	[0.01,0.99]	Estimated from parameter fitting
G	The size of groups in which interactions between agents occur relatively more frequent	[3,50]	Estimated from parameter fitting
γ_G	The probability that a meeting initiated by an agent is with their groupmate	[0.3,0.95]	Estimated from parameter fitting

their next state is R; or the last 2 days of state A, if their next state is I.

2.3 Effect of Traveling

One of the most important feature in the Global-Vax model is its ability in simulating the effect of viral transmission between countries. Cross-country travels spread the disease around the world, hindering regional efforts and necessitating the global coordination of vaccine distribution. Modeling global travels typically requires access to detailed international flight data that include number of flights between airports, number of available seats and occupancy rates etc. Even if such data were publicly available, moving agents from one ABM to another poses significant computation complexity.

Therefore, we decide to simplify the simulation of traveling by only simulating the effect of traveling, not traveling itself, reducing the task to the estimation of the number agents infected by an imported cases.

In each region, we use $n_{\text{from travel}}$ as the notation for the number agents infected by an imported case. We do not model or account for imported cases, rather, we use the prevalence of neighbor regions to make an estimation for $n_{\text{from travel}}$. Here prevalence P is defined as $I_{\text{new}}/N_{\text{pop}}$. $n_{\text{from travel}}$ is proportional to the risk of importation. For region j, the risk is estimated by a weighted sum of the prevalence of other regions:

$$\text{Risk}_j = \sum_{i \neq j} w_{ij} P_i \tag{1}$$

where i stands for all other regions. P_i is the prevalence of region i. w_{ij} represents the effective distance between region i and j. Internationally, this is defined as the inverse of the minimal duration of the flights between the capitals of two countries[1]. Sub-nationally only for the U.S., it is the geographic distance between the capitals of two states. w_{ij} is also scaled by the population size of region i.

To convert the risk into $n_{\text{from travel}}$, we multiply it by a constant, so that per one million, $1 \lesssim n_{\text{from travel}} \lesssim 10$. This constant is tuned during parameter fitting.

Note that when data is available, i.e., the model is not projecting forward into the future, we can use the observed prevalence as P_i to simulate the effect of traveling. This allows us to model interconnectivity between regions while keeping the regional simulations to be independent in practice.

On the other hand, during projection, the implementation of traveling is done by the **Meta** object. When traveling is turned on, at the beginning of each step, $n_{\text{from travel}}$ is calculated for each regional **ABM** object, based on the prevalence of other regions from the last time step. Then $n_{\text{from travel}}$ agents will be randomly chosen and their state will be updated to A. Afterwards, interactions between agents and their state evolution occur only within each individual **ABM**.

For certain regions, I_{new} and $n_{\text{from travel}}$ are less than 1 per million daily. We use an addition parameter p_t to scale down the effect traveling, by drawing a random variable from Bernoulli(p_t) and updating chosen agents to state A only if this variable is 1.

[1]Flight data is searched at `https://tequila-api.kiwi.com`, between 5/1/2021 and 12/31/2021

2.4 Disease Dynamics

The disease dynamics in Global-Vax is based on the SIR model with 5 possible states as described in Section 2.1 and Figure 1.

At each time step t, an agent i moves forward to the state of next time step $a_i(t+1)$, based on its state at current time $a_i(t)$ and some state-dependent rules.

2.4.1 S→A

If a susceptible agent i is contacted by an infectious agent j, there is a probability $P_{\mathrm{S}\to\mathrm{A}}(x_i)$ that i becomes asymptomatically infected, with $P_{\mathrm{S}\to\mathrm{A}}(x_i) = P^0_{\mathrm{S}\to\mathrm{A}} \times S(x_i)$, where $P^0_{\mathrm{S}\to\mathrm{A}}$ is the transmission probability per contact of senior females and $S(x_i)$ is a scale factor, depending on the demographic group x_i. $S(x_i)$ is to be estimated from data.

2.4.2 A→I or R

If agent i becomes asymptomatically infected, the time step that it transitions from S to A is denoted as $T_{\mathrm{S}\to\mathrm{A}}$. At $T_{\mathrm{S}\to\mathrm{A}}$, a random number z is drawn from Bernoulli($P_{\mathrm{A}\to\mathrm{I}}$), if $z = 1$, it is determined that agent i will eventually become I, symptomatically infected, otherwise, the agent will enter R, recovered. Then the algorithm determines how long agent i will stay at A. We denote T_A as the time agent i spent in state A. T_A is drawn from a log-normal distribution, whose mean and variance depends on whether agent i will turn into I or R next. Mathematically, it is expressed as the following:

$$T_A = \begin{cases} \text{Lognormal}(\mu_{\mathrm{A}\to\mathrm{I}}, \sigma^2_{\mathrm{A}\to\mathrm{I}}), & \text{if Bernoulli}(P_{\mathrm{A}\to\mathrm{I}}) = 1 \\ \text{Lognormal}(\mu_{\mathrm{A}\to\mathrm{R}}, \sigma^2_{\mathrm{A}\to\mathrm{R}}), & \text{if Bernoulli}(P_{\mathrm{A}\to\mathrm{I}}) = 0 \end{cases} \tag{2}$$

Agents in state A can be infectious, as mentioned in Section 2.2.1. T_I, the day an agent of state A become infections depends on its next state, as follows:

$$\tau_A = \begin{cases} \text{Lognormal}(\mu_{\mathrm{A}\to\mathrm{I}}, \sigma^2_{\mathrm{A}\to\mathrm{I}})) - 2, & \text{if Bernoulli}(P_{\mathrm{A}\to\mathrm{I}}) = 1 \\ \texttt{Max}(3, \text{Lognormal}(\mu_{\mathrm{A}\to\mathrm{R}}, \sigma^2_{\mathrm{A}\to\mathrm{R}})), & \text{if Bernoulli}(P_{\mathrm{A}\to\mathrm{I}}) = 0 \end{cases} \tag{3}$$

2.4.3 I→R or D

If agent i becomes infected, the time step that it transitions from A to I is denoted as $T_{\mathrm{A}\to\mathrm{I}}$. At $T_{\mathrm{A}\to\mathrm{I}}$, a random number z is drawn from Bernoulli($P_{\mathrm{I}\to\mathrm{D}}$), if $z = 1$, it is determined that agent i will die in the end of its infection, otherwise, the agent will recover. Then the algorithm determines how long agent i will stay at I. We denote T_I as the time agent i spent in state I. If Bernoulli($P_{\mathrm{I}\to\mathrm{D}}$)) = 1, T_I is drawn from Lognormal($\mu_{\mathrm{A}\to\mathrm{I}}, \sigma^2_{\mathrm{A}\to\mathrm{I}}$), with a cutoff at 21 days. Therefore $T_I = \max(\text{Lognormal}(\mu_{\mathrm{A}\to\mathrm{I}}, \sigma^2_{\mathrm{A}\to\mathrm{I}}), 21)$. Otherwise, if If Bernoulli($P_{\mathrm{I}\to\mathrm{D}}$)) = 0, T_I is fixed at 21 days. So T_I can be expressed as:

$$T_I = \begin{cases} \max(\text{Lognormal}(\mu_{\mathrm{I}\to\mathrm{D}}, \sigma^2_{\mathrm{I}\to\mathrm{D}}), 21), & \text{if Bernoulli}(P_{\mathrm{I}\to\mathrm{D}}) = 1 \\ 21, & \text{if Bernoulli}(P_{\mathrm{I}\to\mathrm{D}}) = 0 \end{cases} \tag{4}$$

$P_{\mathrm{I}\to\mathrm{D}}$ is the case fatality of the disease. It depends on the demographic group x_i and is to be estimated from data.

2.4.4 Final States

An agent's state won't change once it reached R or D.

2.5 Agent Interactions

2.5.1 Meetings

A random network is generated at each step to model agent interactions. Intuitively, meetings are restrict to agents within the same region, i.e. the same **ABM** object. This means that potentially any agent can come into contact with anyone else in the same region. As mentioned in Section 2.2.1, π_i denotes the meeting prevalence of agent i. For the implementation of the model, it is defined as the number of meeting an agent will initiate daily. Moreover, π_i also denote the sampling probability (before normalization) of agent i to be chosen by others, since by definition, an agent with higher meeting prevalence is also more likely to be contacted by others. Furthermore, we define an effective meeting as one where viral transmission can potentially occur. Apparently, an effective meeting can only be between an infectious agent and a susceptible one. Even in a worldwide pandemic, the number of people who are contagious at a given time is only a small fraction of the total population. If we impose a restriction in the model that only infectious agents can initiate meetings, we can reduce unnecessary operations significantly. It can be mathematically proven that the average number of effective meetings will be the same for the following two cases: 1) every agent initiate on average K meetings daily; 2) every infectious agent initiate an average of $2K$ meetings daily (See Supplemental Materials). Therefore, for the simulation of disease transmission, rescaling π_i by a factor of 2 and allowing only infectious agents to initiate meetings will be equivalent as allowing all agents to initiate meet π_i others in every step. Since π_i is a parameter to be fitted to data during model calibration, we can omit the factor of 2 within the model all together.

Additionally, we introduce an additional parameter $\gamma_M \in (0, 1)$ as the scaling factor for π_i for agents that are in state I, since it is likely that an agent who is sick would be less inclined to initiate meeting with others.

To sum up, $\hat{\pi_i}$ the number of meeting agent i initiates per step is as follows

$$\hat{\pi_i} = \begin{cases} \pi_i, & a_i = 2 \text{ and } \mathcal{I}_i = 1 \\ \pi_i \cdot \gamma_M, & a_i = 3 \\ 0, & a_i = 1, 4, 5 \end{cases} \tag{5}$$

In the algorithm, meetings are conducted as follows: at step t, for an agent i, π_i agents are sampled with replacement for meeting from the rest of agents $j \neq i$, each with sampling probability $\tilde{\pi}_j$ to be specified in the following text.

2.5.2 Groups Inside a Region

We define groups inside a region to be a collection of agents whose interactions are more frequent between one another of the same group than between agents of different groups. The meeting preference of an agent j viewed by agents in the same group is j's original

meeting preference π_j. On the other hand, j's meeting preference is rescaled by a ratio $\gamma_G < 1$, if viewed by agents outside its group.

We use the symbol a_i^G to denote the group where an agent i is. Then an agent j's normalized meeting preference perceived by another agent i is

$$\tilde{\pi}_j(t) = \frac{\hat{\pi_j} \cdot \Gamma_{ij}}{\sum_{j' \neq i} \hat{\pi_{j'}} \cdot \Gamma_{ij'}.} \tag{6}$$

where

$$\Gamma_{ij} = \begin{cases} 1 & \text{if } a_i^G = a_j^G \\ \gamma_G & \text{else,} \end{cases} \tag{7}$$

with γ_G to be calibrated before initializing the model.

When a model is defined, the group number a_i^G is assigned to each agent. Group size G is an input parameter governing the number of agents in a group. Then the number of groups in a region R is $\lfloor \frac{N_{\text{pop}}}{G} \rfloor + 1$.

When an agent i travels to a new region, a_i^G will be updated and randomly assigned to a group in the new region.

With γ_G in both the numerator and the denominator of Eq. 6, roughly speaking, its effect is to scale down $\tilde{\pi}_j$ by $10\gamma_G$, if there are 10 groups in the region.

2.5.3 Implementation Improvements

To implement the sampling process for meetings, we use `Numpy.random.choice` function in Python. The time complexity of uniformly sampling m times from a list N_{pop} is $O(m)$. However, it is $O(N_{\text{pop}} + m \log N_{\text{pop}})$ for weighted sampling[2], which is the case here for the meeting preference π_i is agent-dependent. This will lead to an overall complexity of at most $O(N_{\text{pop}}^2)$ for each iteration of **ABM** (for the simplicity of explanation, we assume the simulated number of infections agents per step is proportional to N_{pop}). In real time, this is impractical as it means running an **ABM** object with one million agents for the complete evolution of an outbreak would take $\gtrsim$ one week.

To reduce the time complexity significantly, we make use what we call the mixing array. For each time step an infections agent i will meet $\hat{\pi}_i$ agents. Agents to be met are sampled uniformly with replacement from a array of agent IDs, in which an ID j repeats $\hat{\pi}_j$ times. We call this array the mixing array. The mixing array for agents in the group G of region R is labeled $M_{\{R,G\}}$, used in same group sampling, and the array for agents in region R is M_R.

The parameter γ_G is still used to distinguish between meeting with agents from the same group and from other groups. Only now it is the probability that a meeting is between agents in the same group.

The algorithm is as follows:

1. For each agent i, draw a random variable $v \sim \text{Binomial}(\hat{\pi}_i, \gamma_G)$, as the number of meetings with agents in the same group as i.

2. Uniformly draw v samples from $M_{\{R_i,G_i\}}$ with replacement and conduct meetings.

[2] $O(N_{\text{pop}})$ for generate an array to store cumulative sum of weights up to a certain index, and $O(m \log N_{\text{pop}})$ for performing binary search for the indices that correspond to a random number between $[0, N_{\text{pop}})$.

3. Uniformly draw $(\hat{\pi}_i - v)$ samples from M_{R_i} with replacement and conduct meetings.

The probability of drawing j at one sampling is

$$\tilde{\pi}_j(t) = \gamma_G \frac{\hat{\pi_j} \cdot \mathbb{1}(G_i(t) = G_j(t))}{\sum_{j' \neq i} \hat{\pi_{j'}} \cdot \mathbb{1}(, G_i(t) = G'_j(t)} + (1 - \gamma_G) \frac{\hat{\pi_j}}{\sum_{j' \neq i} \hat{\pi_{j'}}} \tag{8}$$

$M_{\{R,G\}}$ and M_R are updated at each step before meetings to account for changes due to the travelling of agents and their transitions from one state to the next. To avoid redundancy, the group level mixing arrays will only be updated if agents with the corresponding groups have changed their mixing preference.

Now, the time complexity of an iteration, spent on the sampling alone, reduces to $O(N_{\text{pop}})$. Updating $M_{\{R,G\}}$ and M_R will at most takes additionally $O(N_{\text{pop}})$ time. So the overall time complexity for an iteration is still only $O(N_{\text{pop}})$.

2.6 Vaccination

Given a vaccine plan detailing the vaccine efficacy, when and how many agents will be vaccinated in a time step, the **ABM** object can simulate the effect of agents being vaccinated accordingly. In our model, the vaccine efficacy $\mathcal{E}$ is translated into the reduction in risk of a vaccinated agent getting infected in the future. That means if the agent's state is S when they are given the vaccine, $P_{\text{S}\to\text{A}}$, their probability of becoming infected when coming into contact with an infectious agent, will be reduced to $P_{\text{S}\to\text{A}} \times (1 - \mathcal{E})$ in all future steps. For the case of COVID-19, any person who is not dead or currently sick is eligible for COVID-19 vaccine. So in our model, we randomly sample from the pool of eligible agents, whose state are in S, A or R, to be vaccinated at each step. The number of agents to be sampled is the number of available vaccines in the vaccine plan. If the chosen agent is in state A, they will be moved to state R with probability $\mathcal{E}$. If they are in state R, then nothing will be done. The vaccination plan will stop, once there is no more eligible agents at a given time step. In addition, we allows for the option to impose priority on senior vaccinations, meaning the above algorithm will first choose from eligible seniors, when there are no more seniors to be vaccinated, all other eligible agents will be chosen.

3 Calibration and Validation

As the Globa-Vax model is designed for the simulation of global vaccine allocation, this section provides a demonstration for how it can be calibrated and parameterized so that the model prediction fits actual data. However, the exploration of different vaccine allocation scenarios is beyond the scope of this paper. For an example of using the Globa-Vax model to optimize COVID-19 vaccine allocation, see [6].

For the calibration, we choose to include 148 countries which accounts for more than 90% of the global population and over 95% of COVID-19 cases in the world. We model one country with one **ABM** object, except for the U.S., where we use one **ABM** per state. For each **ABM** object, we include one million agents. Holding the number of agents fixed for all regions means one less variable to fit for during calibration. And we use the scaling factor S_{pop} to re-scale model predictions in order to match observed data.

3.1 Input Data

3.1.1 Demographic Characteristics

Our simulated population should preserve demographic characteristics that the actual COVID data has shown, including:

1. The fraction of different demographic groups in the population.
2. The confirmed case fraction by sex and age.
3. The case fatality by sex and age.

Once we have the data above, we can feed them into our models. The case fraction need to be first translated into a sex and age dependent factor $S(x_i)$, that will scale $P_{\mathrm{S}\to\mathrm{A}}$ to affect the probability of an agent in a specific sex and age group getting infected when contacted by an infectious agent. This will result in sex and age fractions in the simulated cases reflecting the actual demographic fractions. The scale factor for a demographic group can be computed by dividing its case fraction by its population fraction.

Note, for U.S., race-disaggregated data is also available[3]. So for U.S. only, we also model the racial demographic characteristics.

Demographic compositions for all countries can be obtained from the United Nations' world population prospect data [4]. However, only a few countries have public available sex and age-disaggregated COVID case and death data. We proceed by making approximations for the countries where data is not available, with the case proportion and case fatality from its nearby neighbors where data is available.

It is still not feasible to look for all the data from 148 countries. We therefore only focus on countries that make up 90% of the cases in the world[5]. There are 49 countries that make the list, out of which, we are able to obtain case demographic data from 19 countries.

We then group all countries into WHO regions[6]. The case fractions and case fatalities of countries in a WHO region are the average of the values of the countries in that region where case demographic data is available.

3.1.2 Vaccine Dosage and Types

For the vaccine plan, we estimate the daily dosage per million for each region from GOVEX data[7]. We take the average of daily dosages over a period of time and apply that to our models. We assume two types of vaccines, one with efficacy 65.8% (representing J&J) and the other with 95% (representing two-dosage vaccines such as Pfizer and Moderna). For

[3]Population estimates are available at `https://www.census.gov/newsroom/press-kits/2020/population-estimates-detailed.html`. COVID-19 case data with sex, age and race information is available at `https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf`.

[4]`https://population.un.org/wpp/Download/Standard/Population/`

[5]Estimates come from the cumulative case counts as of 4/16/2021 at `https://github.com/CSSEGISandData/COVID-19`

[6]defined in `https://en.wikipedia.org/wiki/List_of_WHO_regions`

[7]`https://github.com/govex/COVID-19/tree/master/data_tables/vaccine_data/global_data`

simplicity, we assume each type of vaccine accounts for half of all vaccinations and they start taking effect right away. We also prioritize seniors for vaccination, as is consistent with the vaccination order in most countries.

3.2 Parameter Fitting

To obtain predictive power from our models, first we need to find the set of parameters that provide a good fit to the cumulative case time series in the past. We then keep on evolving the models and allow 35 days for validation, checking that the models continue to agree with actual data within an acceptable margin of error. From this point on, we can potentially apply different vaccine donation strategies to the models and test how effective each of the strategies is at reducing COVID cases.

Note that running **ABM** objects with one million agents each takes ∼10 min of time (for ∼300 time steps) and 450 MB of memory. We try to evolve regional models as independently from each other as possible, so we can run them in parallel. Only when performing validations and projections do we introduce the **Meta** object to handle traveling and run **ABM** objects in series.

We fit our models to data at the time frame 1/1/2020-6/30/2021. The fitting is divided into two periods, followed by a validation period.

1. In the first part the fitting is performed on disconnected regional models with the same default demographic characteristics. It is for fitting the model for parameters that are significant at governing the shape and size of the outbreak curve.

2. In part two, traveling, vaccination and regional demographic characteristics are added to the models, and to account for changes in the modelled curves that are introduced, parameters are slightly adjusted to ensure a good fit to the data curves. For most of the regions, the two periods are divided at 5/31/2021.

3. To validate the models, we look for out-of-sample agreement between observed data and the model prediction over the period of 7/1/2021 to 8/14/2021, for 35 days. At this stage, we join all the **ABM** objects in one **Meta** model. Interconnections are realized by using simulated regional COVID prevalence to compute the risk of importation (Equation 1).

3.2.1 First Fitting Period

During the first period, regions evolves independently within individual **ABM**s.

We do a grid search with 6300 combinations of the values of four parameters: $\{G,\ \gamma_G,\ \gamma_M,\ P_{\mathrm{S}\to\mathrm{A}}\}$, along with 10 different seeds for the *Numpy* random number generator, which means 63000 simulations in total are searched. All these simulations have the same number of population size: one million. The initial condition is set to have 122 agents in state A and 100 in state I. The maximum number of agents one can meet daily is set to be 30. Fractions of different demographic group are set to the same default values (otherwise the grid search would need to be repeated for regions with different demographics).

Our model is based on simple SIR models, where outbreaks start at time step 0 and there is only one wave. In reality, the COVID outbreak started at different time in each region,

and some regions have experienced several waves of outbreak until now. So we make the simplifying assumption that most regions have only one wave and we start each regional **ABM** on the day when its cumulative COVID cases reached 100 per million. This way, each **ABM** will have evolved for different number of days on 5/31/2021. For the regions that have experienced distinct waves of outbreak, we start their simulations at around the beginning of their last wave, with the same initial condition as mentioned above. So to obtain predictions from simulations, we will add the actual cumulative cases at the start day to offset the difference between data and simulation. We use a global day counter in all **ABM** objects to ensure the correct day counts.

The end days of the first periods vary among regions. As we intend not to include vaccination during this period (since the inclusion would open up another dimension in parameter exploration), the end day should be when residents at a certain region started to receive vaccination, or 5/31, whichever comes first. We define the first day of vaccination in the model to be when the actual vaccination coverage reached 3% of the region's population. Thus, if a region's vaccination reached 3% before 5/31, say 3/20, we will evolve its **ABM** until 3/20, without vaccination and traveling, after that it enters its second fitting period (will be described at next section). We denote this day as $d_{3\%}$.

The best-fit parameters are defined as the set of $\{G,\ \gamma_G,\ \gamma_M,\ P_{\mathrm{S}\to\mathrm{A}}\}$ and seed number that gives the best agreement/lowest RMSE in the last 15 days of the first fitting period between the actual cumulative cases and the simulated results from **ABM**. This is because the agreement with recent data in terms of outbreak size and curvature is what matters for projection forward.

To check for the in-sample goodness of fit, we compute the mean absolute percentage error (MAPE) for the last 15 days of the first fitting period. We achieve an excellent fit, with 95% of the 198 regions having MAPE less that 2%.

3.2.2 Second Fitting Period

For a region that reached a 3% population coverage in vaccination before 5/31, we vaccinate 3% of its **ABM** agents on $d_{3\%}$. Entering the second fitting period, until 5/31, the number of agents vaccinated per step is the average daily dosage per million between $d_{3\%}$ and 5/31 from GOVEX data. Between 6/1 and 6/30, for all regions, with or without reaching 3% by 5/31, the number of agents being vaccinated will also be estimated from GOVEX as the average daily dosage per million during this period. The last estimate will also be used as the daily dosage projecting forward.

We also add the effect of traveling by calculating $n_{\text{from travel}}$. We use the regional prevalence from actual data, instead of from simulations, to avoid having to introduce the **Meta** object at this point, so each regional **ABM** can still run independently and in parallel. We assume that at most only 2.5% of daily new cases come from contacts to an infectious imported case. So we set the travel scaling down probability $p_t = \min(1, \text{average daily new cases} * 0.025)$, where average daily new cases is estimated in the period of 5/11-5/31.

At this stage, we also modify the demographic characteristics in the **ABM** objects, so that the case fraction, case fatality and the overall fraction of each demographic group are consistent with those of the specific regions. A new scale factor for $\mathrm{S} \to \mathrm{A}$ will be computed.

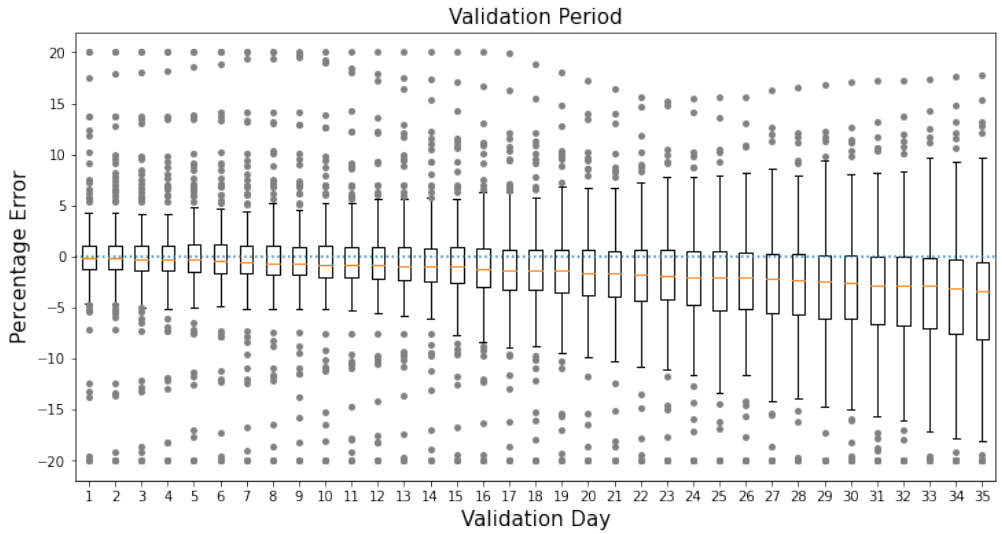

Figure 2: The distribution of percentage error over the validation period.

And we maintain the same population average $P_{\mathrm{S}\to\mathrm{A}}$ by doing the following calculation:

$$P^{\mathrm{new}}_{\mathrm{S}\to\mathrm{A}} = P^{\mathrm{old}}_{\mathrm{S}\to\mathrm{A}} \times \frac{\sum_d S(d)^{\mathrm{new}} \times \mathrm{new\_population\_fraction}[d]}{\sum_d S(d)^{\mathrm{old}} \times \mathrm{old\_population\_fraction}[d]} \tag{9}$$

where d stands for a demographic group.

Even if we have converted $P_{\mathrm{S}\to\mathrm{A}}$ according to new demographic characteristics, adding vaccination and traveling to the model may still have an effect on the pandemic curve. Here we make a trade-off between time complexity and goodness of fit by only only conducting a refit for the parameter $P_{\mathrm{S}\to\mathrm{A}}$. We do a grid search by multiplying $P_{\mathrm{S}\to\mathrm{A}}$ with a factor from an array range from 0.1 to 1.9 with a step size of 0.1. For each region, the best-fit model is the one with the lowest RMSE between data and prediction of the second period. The MAPE of this period has larger variance and is not as good as the first period. Although 90% of the regions achieve a MAPE less than 7%, the 95% quantile of MAPE is at 15%. This is a result of the many new features included in the **ABM** objects at the second period, as well as the data itself becoming more complicated as vaccination has started in many regions, making the disease transmission more complex.

3.3 Results

To validate the models, we first join all 198 **ABM** objects in one **Meta** model. As the goal is to check how the model performs on its own, We stop using the observed prevalence to feed in Equation (1) to simulate effect of travel. Rather, we using simulated regional prevalence. That means **ABM** objects are interdependent on one another. With all other settings the same as the second fitting period, we project the **Meta** model forward over the period of 7/1/2021 to 8/14/2021, for 35 days.

Figure 2 shows the five-number summaries of the out-of-sample percentage difference between observed data and the regional model predictions over the validation period. Overall,

the Global-Vax model has made reliable forecasts. Most of the regional predictions deviate from the observed data by less than 10%. However, there is a clear trend of underestimation, as time goes on. This is mainly due to occurrences of new outbreaks during this period. Although the Global-Vax model is designed and calibrated to fit transmission dynamics, it is not yet able model or predict unforeseen changes in government intervention or other human behaviours, which are likely what causes some of the recent COVID surges. A better forecast may be achieved by adding a time dependent scaling factor to the transmission rate, to simulate, for example, the strengthening or easing of interventions. In addition, it may also help to include feedback loops so that the level of future intervention depends on the current prevalence. However, these possible improvements can significantly increase the complexity of our model, and given the overall satisfying accuracy achieved so far, they are left to future research.

4 Discussion

The severity of COVID-19 pandemic and limited supply of global vaccines have demanded for computational models that can provide accurate forecast and explore the effectiveness of different vaccine allocation scenarios. To this end, we have built a mathematical model to simulate viral transmission on the global scale, allowing for effects of international travels, regional demographics and vaccination. The model can project forward in time and forecast future cases. So we can learn where to allocate vaccines for maximum impact. To sum up, the main contributions of this paper are as follows:

- We built one of the first global-scale agent-based model for COVID transmission.
- The model allows for flexible experimentation for custom vaccine allocation strategies.
- We developed a fast, linear time implementation for agent interactions.
- The Global-Vax model has been thoroughly validated using real-world pandemic, demographic, and travel data.

Despite its contributions, the Global-Vax model is not without limitations.

We have made some simplifying assumptions in the model, such as using a random contact network for agent interaction, only accounting for vaccine effectiveness after the second (or the last) dose, assuming homogeneity in vaccine response. Although these assumptions have not impact significantly our model performance, prediction accuracy would no doubt benefit from including more complex and realistic elements in the model, for example, contact networks that consists of social structures and hierarchies, heterogeneous vaccines response that depends on agent properties and so on.

The current implementation of traveling assumes constant effective distance between regions, although this assumption is not unreasonable for short term projections. In addition, simulating effects of imported cases without those exported cases results in a net increase in global cases at each step, however, the bias it creates is negligible compared to the recent surges in observed cases. With increased availability and reliability of both COVID-19 data

and global travel data, future efforts will focus on improving modeling on mobility network and its role in disease dynamics.

At the moment, the 198-region **Meta** model takes around 4 hours to run a 3-month projection. The time complexity may be improved by implementation of multiprocessing. This would allow for longer term projection, which make it feasible to better explore the effects of different vaccine allocation scenarios.

Author contributions

QL and YH conceptualized the study, collated the data, developed the models, visualized the results, and wrote the manuscript. Both authors had full access to all the data in the study.

Conflict of Interest

The authors declare no conflict of interest.

Data sharing The code and data for this study are available on the first author's GitHub page.

Acknowledgments

The study was funded by the Bill and Melinda Gates Foundation. The simulations and computations of the project were conducted using scientific computing services at the Maryland Advanced Research Computer Center (MARCC).

References

[1] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020;20:533-4.

[2] World Health Organization, WHO coronavirus disease (COVID-19) dashboard, available at `https://covid19.who.int`. Last accessed online 12/21/2020.

[3] Centers for Disease Control and Prevention, CDC COVID data tracker, available at `https://covid.cdc.gov/covid-data-tracker`. Last accessed online 12/21/2020.

[4] C. C. Kerr, R. M. Stuart, D. Mistry, R. G. Abeysuriya, G. Hart, K. Rosenfeld, P. Selvaraj, R. C. Nunez, B. Hagedorn, L. George, A. Izzo, A. Palmer, D. Delport, C. Bennette, B. Wagner, S. Chang, J. A. Cohen, J. Panovska-Griffiths, M. Jastrzebski, A. P. Oron, E. Wenger, M. Famulare, D. J. Klein, *medRxiv preprint medRxiv:2020.05.10.20097469* **2020**.

[5] D. Brockmann, D. Helbing, Science. 2013 Dec 13;342(6164):1337-42. doi: 10.1126/science.1245200.

[6] Q. Li, Y. Huang, 2021, Optimizing Global COVID-19 Vaccine Allocation: An Agent-based Computational Model of 148 Countries. Manuscript in preparation.

A Appendix

In this section, we will show that for agent interactions, we can achieve the same results asymptotically with only infectious agents to initiate meetings, as requiring all agents to initiate meetings. Note for simplicity we will no distinguish in- and out-group meetings in this section.

A.1 Notations

- N_I: number of infectious persons
- N_S: number of susceptible persons
- N_O: number of persons in other statuses
- N: total number of people; $N = N_I + N_S + N_O$

A.2 Setting One: everyone initiates meeting

When everyone can initiate meetings, person i initiates $X_i \sim \text{Uniform}(0, M_{\max})$ meetings. The total number of meetings in the population is the sum of N independent discrete uniform distributions, $\sum_{i=1}^{N} X_i$. Its expected value and variance are,

$$E\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} E(X_i) = N\frac{M_{\max}}{2} \tag{10}$$

$$var\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} var(X_i) = N\frac{M_{\max}^2}{12} \tag{11}$$

Only meetings between an infectious person and a susceptible person are effective for virus transmission. Choosing two persons from a population of N to meet, the total number of different combinations is $C_N^2 = \frac{N(N-1)}{2}$, each of which has the same possibility. The number of effective combinations is $N_S N_I$. Therefore, the proportion of effective combinations among all possible combinations is $\frac{N_S N_I}{C_N^2}$. The expected total number of effective meetings is

$$E\left(\frac{N_S N_I}{C_N^2}\sum_{i=1}^{N} X_i\right) = N\frac{M_{\max}}{2}\frac{N_S N_I}{C_N^2} = \frac{M_{\max} N_S N_I}{N-1} \tag{12}$$

On average, every infectious person is involved in $\frac{M_{\max} N_S}{N-1}$ effective meetings. Lowering this number is important for stopping the virus spreading.

If we consider the proportion of effective meetings as fixed, then the variance of the number of effective meetings is,

$$var\left(\frac{N_S N_I}{C_N^2}\sum_{i=1}^{N} X_i\right) = \left(\frac{N_S N_I}{C_N^2}\right)^2 \frac{N M_{\max}^2}{12} = \frac{N_S^2 N_I^2 M_{\max}^2}{3N(N-1)^2} \tag{13}$$

For a small N, it is easy to derive the probability mass function using convolution operations for $\sum_{i=1}^{N} X_i$. When N is large enough, the distribution is asymptotically Gaussian.

A.3 Setting Two: only the infectious initiates meetings

When only infectious persons can initiate meetings, infectious person i initiates $Y_i \sim uniform(0, M'_{\max})$ meetings.

$$E(Y_i) = M'_{\max}/2 \tag{14}$$

$$Var(Y_i) = \frac{(M'_{\max})^2}{12} \tag{15}$$

The total number of meetings is a sum of N_I independent uniform distributions. Its expected value and variance of total number of meetings are,

$$E\left(\sum_{i=1}^{N_I} Y_i\right) = N_I E(Y_i)) = N_I \frac{M'_{\max}}{2} \tag{16}$$

$$var\left(\sum_{i=1}^{N_I} Y_i\right) = N_I var(Y_i) = N_I \frac{(M'_{\max})^2}{12} \tag{17}$$

A meeting initiated by an infectious person is effective only when the meeting partner is a susceptible person, which occurs with a probability of $\frac{N_S}{N-1}$. Therefore, the expected number of effective meetings is

$$E\left(\frac{N_S}{N-1}\sum_{i=1}^{N_I} Y_i\right) = N_I \frac{M'_{\max}}{2}\frac{N_S}{N-1} \tag{18}$$

Like above, if we consider the proportion of effective meetings as fixed, then the variance of the number of effective meetings is,

$$var\left(\frac{N_S}{N-1}\sum_{i=1}^{N_I} Y_i\right) = N_I \frac{(M'_{\max})^2}{12}\left(\frac{N_S}{N-1}\right)^2 = \frac{N_I (M'_{\max})^2 N_S^2}{12(N-1)^2} \tag{19}$$

When the number of infectious people N_I is small, we can use convolution operations to derive the distribution of total number of meetings. When N_I is large enough, the distribution is asymptotically Gaussian.

A.4 Compare the two settings

In order for the two settings to have the same expected value of effective meetings, we need

$$N_I M_{\max} \frac{N_S}{N-1} = N_I \frac{M'_{\max}}{2}\frac{N_S}{N-1} \tag{20}$$

i.e., $M'_{\max} = 2M_{\max}$. Then variance ratio of the two settings is

$$\frac{\frac{N_S^2 N_I^2 M_{\max}^2}{3N(N-1)^2}}{\frac{N_I (M'_{\max})^2 N_S^2}{12(N-1)^2}} = \frac{N_I}{N} \tag{21}$$

In sum, the number of effective meetings in the two settings are both asymptotically Gaussian distribution if the population size and the number of infectious persons are large enough. The two distributions have the same expected value, but the second setting has a variance larger than the first one by the proportion of infectious people in the population.

A.5 A simple interpretation of $M'_{\mathbf{max}} = 2M_{\mathbf{max}}$

When everyone can initiate meetings, an effective meeting can be either 1) initiated by an infectious person and accepted by a susceptible person; or 2) initiated by an susceptible person and accepted by an infectious person.

When only infectious persons can initiate meetings, every infectious person needs to initiate **twice** as many meetings in order to generate the same total number of effective meetings as in the current setting.