AI-Driven Agent-Based Models to Study the Role of Vaccine Acceptance in Controlling COVID-19 Spread in the US

Parantapa Bhattacharya*, Dustin Machi*, Jiangzhuo Chen*, Stefan Hoops*, Bryan Lewis*,
Henning Mortveit*, Srinivasan Venkatramanan*, Mandy L. Wilson*, Achla Marathe*,
Przemyslaw Porebski*, Brian Klahn*, Joseph Outten*, Anil Vullikanti*,
Dawen Xie*, Abhijin Adiga*, Shawn Brown†, Christopher Barrett*, Madhav Marathe*,
*University of Virginia, USA
†Pittsburgh Supercomputing Center, USA

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671811

***Abstract*—We study the role of vaccine acceptance in controlling the spread of COVID-19 in the US using AI-driven agent-based models. Our study uses a 288 million node social contact network spanning all 50 US states plus Washington DC, comprised of 3300 counties, with 12.59 billion daily interactions. The highly-resolved agent-based models use realistic information about disease progression, vaccine uptake, production schedules, acceptance trends, prevalence, and social distancing guidelines. Developing a national model at this resolution that is driven by realistic data requires a complex scalable workflow, model calibration, simulation, and analytics components. Our workflow optimizes the total execution time and helps in improving overall human productivity.**

This work develops a pipeline that can execute US-scale models and associated workflows that typically present significant big data challenges. Our results show that, when compared to faster and accelerating vaccinations, slower *vaccination rates* due to vaccine hesitancy cause averted infections to drop from 6.7M to 4.5M, and averted total deaths to drop from 39.4K to 28.2K nationwide. This occurs despite the fact that the final vaccine coverage is the same in both scenarios. Improving vaccine acceptance by 10% in all states increases averted infections from 4.5M to 4.7M (a 4.4% improvement) and total deaths from 28.2K to 29.9K (a 6% increase) nationwide. The analysis also reveals interesting spatio-temporal differences in COVID-19 dynamics as a result of vaccine acceptance. To our knowledge, this is the first national-scale analysis of the effect of vaccine acceptance on the spread of COVID-19, using detailed and realistic agent-based models.

***Index Terms*—High Performance Computing Workflow, Agent-Based Modeling and Simulation, Covid-19 Vaccine Acceptance**

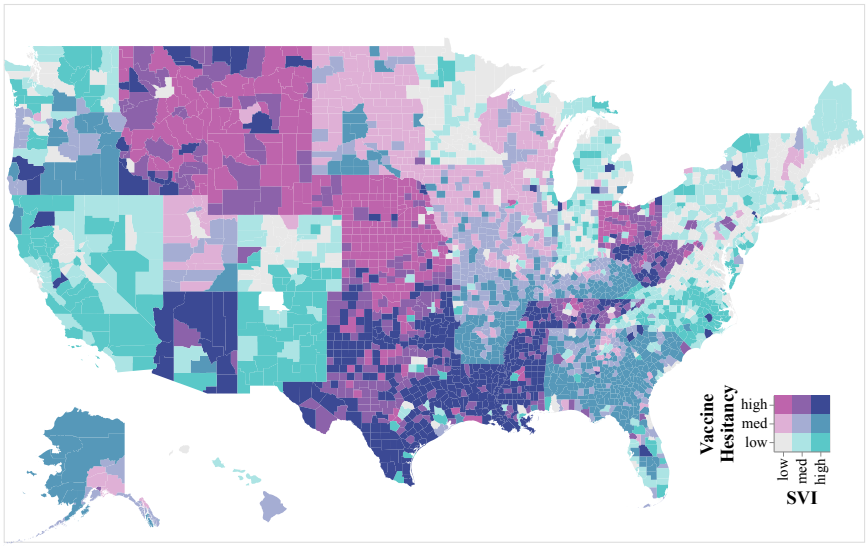
I. Introduction

The COVID-19 pandemic continues to pose significant social, economic, and health challenges. The economic losses have already been in the trillions of dollars just in the United States (US). As of August 2, 2021, this pandemic has caused over 198 million confirmed cases and 4.2 million deaths worldwide. The pandemic caused significant waves in India in April 2021 that resulted in unprecedented mortality and morbidity.

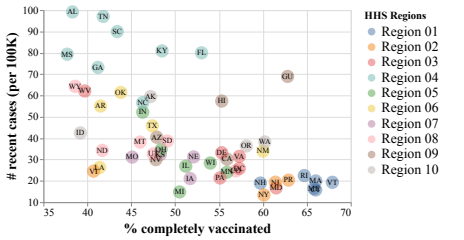
The introduction of COVID-19 vaccines has proven vital in the control of the pandemic in many western countries. The US had seen a significant decline in cases earlier this year as an increasingly larger fraction of the population was vaccinated. Initially, when the vaccine rollout began in January 2021, the availability of vaccines was the biggest bottleneck in the drive to vaccinate millions of Americans, but now states are awash in vaccines. The problem of vaccine allocation in the US has shifted from a supply-side problem to a demand-side problem. In January 2021, there was a limited number of vaccines and many individuals wanting to take them. As of August 2021, we are starting to hit the vaccine hesitancy wall wherein *all* eligible individuals can get a vaccine, but many individuals are unwilling.

Figure 1a presents a bivariate choropleth plot showing the relative distribution of the social vulnerability index [9] and vaccine hesitancy [13] for different counties. Figure 1a suggests vaccine hesitancy has the potential to impact socially vulnerable populations disproportionately. Additionally, Figure 1b and 1c show scatter plots relating the vaccination status and recent case counts across US states and counties in the state of Virginia, respectively. As expected, these figures indicate that states and counties with lower vaccine coverage had higher incidence of disease spikes.

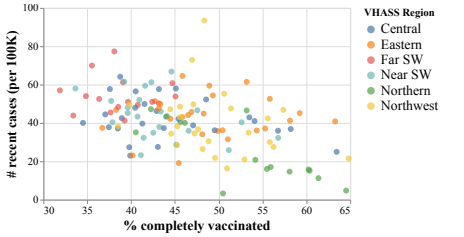
Summary of contributions and significance. In this paper, we study the problem of controlling the COVID-19 pandemic using efficient allocation of vaccines. Three key constraints are modeled: (i) vaccine acceptance, (ii) levels of vaccinated individuals and (iii) numbers of infected individuals. These factors vary substantially across the US, as can be seen in Figure 1. Furthermore, the variation is not just between states, but even within states, making this a challenging problem. Such spatial, temporal and social variabilities are best captured using agent-based models; we describe such a model here – a scalable, high performance, national agent-based modeling framework. We use a realistic digital twin of the US social contact network (288 million nodes, 12.59 billion edges interacting in a time-varying manner). The spatially explicit, time-varying social contact network captures the spatial and temporal heterogeneities. Our basic approach builds on our early work in [19], [15], [25], [3], [33], [27], [12]; this



(a) Social Vulnerability Index vs. Vaccine Hesitancy across US counties.



(b) Recent case counts vs. vaccine coverage across US states.



(c) Recent case counts vs. vaccine coverage across counties in the state of Virginia.

Fig. 1: Fig. 1a shows the relative distribution of the social vulnerability index [9] and vaccine hesitancy for different counties. County-wise vaccine hesitancy is obtained from the Facebook survey data [13] using the dataset as of August 2, 2021. Fig. 1b shows the distribution of the mean number of recent cases (per 100K people) compared to vaccine coverage for the different states in the US as of September 2, 2021. This figure indicates that states with lower vaccine coverage had higher incidence of disease spikes. Fig. 1c shows the distribution of the mean number of recent cases per 100K people compared to vaccine coverage for different counties in the state of Virginia as of September 2. 2021. Localities with higher vaccination coverage, such as Albemarle County and Fairfax City, had lower case rates. With the increase in the spread of the Delta variant of COVID-19, remote and lower vaccinated counties showed increasing case counts.

approach has been validated and used to support H5N1, H1N1 and Ebola response efforts [15], [3]. The need for such AI-driven simulations is also described by Foster et al. in a recent report [17].

In this paper, we make the following advances. *First*, the models are data-driven and use detailed *county-level* information about vaccine production schedules, vaccine acceptance, disease progression parameters, and non-pharmaceutical interventions for the US. The agent-based models had to be extended to represent vaccinated individuals and production schedules. Developing a data integration pipeline that can execute US-scale models and associated workflows presents significant big data challenges (See Section III). *Second*, executing these large-scale simulations requires *high performance computing* (HPC) resources, and availability of such resources is limited; we present a novel approach that uses two separate HPC clusters simultaneously. Workflow orchestration across two large-scale clusters to reduce the overall execution time and improve human productivity is one of the central contributions of this paper. *Finally*, we carry out a detailed analysis of the vaccine allocation problem. We are interested in understanding the spatial and temporal heterogeneity across the US states. *To the best of our knowledge, this is the first time US-scale models with realistic social contact networks have been used to study the impact of vaccines and vaccine acceptance.* In [32], Perrault et al. highlighted the need for developing operational AI systems, calling it a data-to-deployment pipeline. Our work takes a step toward making computations seamless across multiple HPC systems, and demonstrates how HPC clouds can be used for real-world problems in public health.

Insights into the pandemic response. Analysis of the simulations yield interesting insights; we summarize them briefly here. Further details can be found in Sections VI and VII.

1. Spatial and network heterogeneity matters: Effectiveness of vaccine acceptance differs in different states. Averted infections and deaths per 100K individuals can be as high as over 6000 and 40 respectively in some states, and as low as below 1000 and 5 in others. This is a 5–7 fold variation. Interestingly, state population size, instead of vaccine acceptance level, seems to play a more important role.

2. Role of vaccine hesitancy: Vaccine hesitancy reduces the effectiveness of vaccinations. Slower vaccination rates due to saturation from vaccine hesitancy causes total infection aversion to drop to 4.5M (from 6.7M in the fast vaccination case where vaccination rates keep increasing) and total death aversion to drop to 28.2K (from 39.4K in the fast vaccination case) nationwide. This occurs despite the fact that the final vaccine coverage is the same in both scenarios, only the speed of vaccination is different. This demonstrates the significant impact of vaccine hesitancy, and suggests the need for focused efforts to increase acceptance quickly; this has already begun in several states which have started to offer lotteries.

3. Infections and deaths can be further reduced by improving vaccine acceptance. Improving vaccine acceptance by 10% in all states can improve total infection aversion from 4.5M to

4.7M (a 4.4% improvement) and total death aversion from 28.2K to 29.9K (a 6% increase) nationwide. This provides a significant improvement given the currently assumed levels of acceptance.

An extended version of this paper with additional appendices can be found at [1].

II. Related Work

Prioritizing and optimizing vaccine allocation for various epidemiological objectives has been an active research area in infectious disease modeling for over a decade [29], [28], [39]. While various approaches have been studied [23], [40], [31], [22], [7], [18], age-based vaccine prioritization remains the most commonly used [16], [18], though the exact ratios vary based on infection and severity risk of the disease. For COVID-19, there have been multiple attempts at understanding the utility of age- and serostatus-based approaches [6], spatial heterogeneity [24], and global allocations [21]. In the context of the US, the importance of maintaining moderate levels of non-pharmaceutical interventions while targeting high vaccination uptake was highlighted in a multi-model effort in [4].

Most of the approaches above, however, are limited in model fidelity, and do not explicitly capture the heterogenous network structures and detailed interventions that might be in place during a vaccination campaign. Recently, [11] studied the utility of contact structure-based vaccination strategies using a detailed agent-based modeling approach for the state of Virginia. Simpler agent-based representations have also been used to study the impact of delaying the second dose [30]. While some recent studies ([34], [14]) have attempted to address the question of vaccine demand constrants, to the best of our knowledge this is the first study conducted on a national scale using detailed agent-based models and realistic datasets on vaccine rollout and acceptance.

In an previous study [27], we had presented a different pipeline from what is described in this study. Compared to [27], which scheduled tasks statically (apriori), a major innovation of this work is the use of dynamic scheduling of tasks/jobs which significantly improves task throughput. Additionally, the pipeline presented in this work also saves significant resources by bringing up/down database servers (serving the population data) dynamically. In contrast, in our earlier work [27] all the database servers were started in the beginning. Finally, in a very recent study [2] we improve on the present pipeline further. The new pipeline comprises a number of enhancements that further reduce the overall running time and make the pipeline fault-tolerant. Additionally it has a new service that automates task distribution on the two clusters – the present work does this task manually.

III. End-to-end vaccine study workflow

Vaccine study refers to the study of epidemiological outcomes arising from various vaccination allocation and prioritization strategies under different assumptions about vaccine acceptance and vaccine administration schedules. Such studies usually consist of a factorial experiment design where a small

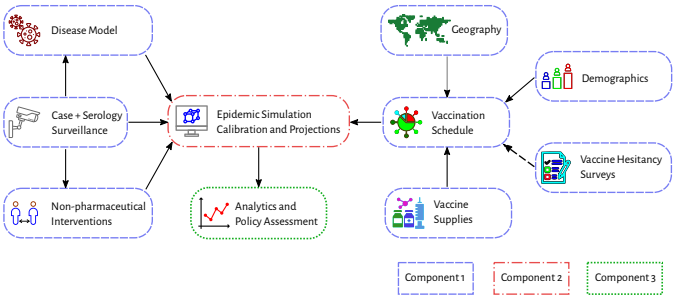


Fig. 2: Overview of the different components of the workflow pipeline as well as their inputs and outputs.

set of critical parameters about the scenario are varied to understand their effects on the epidemiological outcomes. The specific vaccine study discussed in this paper includes 8 configurations (cells), and its experimental settings are described in Section V.

Designing vaccine studies that incorporate all available information in a timely manner generates a number of BigData-related challenges.

Figure 2 shows the multitude of data sources that are used in our workflow to generate accurate and realistic scenario configurations. Incorporating data from so many sources leads to 'volume', 'velocity', 'veracity' and 'variety' related BigData challenges. The vaccine study workflow used here for studying the impact of vaccination on the COVID-19 pandemic, under various assumptions of allocation/prioritization, acceptance, schedule, and efficacy uses the EpiHiper simulator. The confirmed case count data used for the workflow includes data for over 3000 US counties for over 350 days. The factorial design used in this paper needed over 6120 simulation instances: 2 vaccine acceptance levels $\times$ 2 vaccination schedules $\times$ 2 prioritization schemes $\times$ 51 states $\times$ 15 replicates. The individual-level output data, which includes 8 cells $\times$ 51 states $\times$ 15 replicates $\times$ multiple millions of state transitions, contains multiple billion entries and is over 2TB in size. The synthetic population data used for the simualtions includes person, household, location, and activities data. Our analysis mainly uses the person data, which is over 20GB in size. The size of the aggregate epidemiological data, consisting of 8 cells $\times$ 51 states $\times$ 15 replicates $\times$ 300 days $\times$ 180 health states $\times$ 3 counts (number of nodes entering a state, number of nodes currently in a state, and cumulative number of nodes entering a state), contains over a 1 billion entries and is over 2.5GB in size.

Component 1 collects surveillance data on daily confirmed case counts at the county level for over 3000 counties in the US. It also collects and synthesizes county-level data on vaccine acceptance, allocation, prioritization, and vaccination schedules. Once the disease model is ready for each state, we initialize the simulation configurations by combining the disease models with non-pharmaceutical interventions, and incorporating information about prior and recent infections. See Section V for more information.

Component 2 starts by calibrating EpiHiper simulations to

match the estimated *effective reproduction number* at the state level (see Section V-C for more details). This step is executed on our home cluster (Section IV-C). Once the simulation configurations related to the vaccine study have been created, they are then sent to the more powerful remote cluster. As stated above, component 2 requires running 6120 HPC simulations. Even when using the larger remote HPC cluster, executing such a large number of HPC jobs requires additional tooling to run these simulations efficiently.

For this purpose, we had to develop a custom scheduler that is aware of the semantics of our EpiHiper simulations (see Section IV for more details). Since the remote HPC cluster is a shared resource, different scheduling algorithms can be applied to run the simulations depending on the availability of computing resources.

Component 3 summarizes simulation output data and transfers the summary data to the home cluster. Executing such a large number of simulations generates terabytes of data, and creates 'volume' and 'velocity' related BigData challenges. To manage the complexity that arises from handling such large volumes of simulation output data, we compute summarized results (a few gigabytes in size) which are then copied back to our home cluster for analysis. Here, it is augmented with demographic, socioeconomic, and network attributes of individuals from our synthetic population data. Next, we extract aggregate-level spatio-temporal epidemiological measures, e.g. state-level daily infections and deaths in each age group, for further analysis. In this study, we join the data of averted infections and deaths with the synthetic data, as well as vaccine acceptance data, to derive public health policy implications. This step represents the 'value' generation component of the overall BigData pipeline presented in this paper. See Sections V, VI and VII for discussion related to this component.

The end-to-end workflow and the corresponding study design presented in this paper are very flexible and can be updated as input data, such as vaccine acceptance rates or disease prevalence, change over time, or when we change the objective questions to investigate.

IV. Workflow Orchestration for Component 2

In this section, we describe details of component 2. Figure 2 shows an overview of the overall inputs and outputs of the simulation models.

A. Simulation models

The epidemic simulations presented in this paper make use of EpiHiper, an agent-based discrete time simulation platform for infectious disease spread in a social contact network. EpiHiper is implemented as a distributed memory program written in C++/MPI. EpiHiper computes probabilistic disease transmission between nodes (representing individuals) in a network of edges (representing interactions between individuals), as well as the disease progression within each infected individual. It relies on synthetic populations, which are made accessible to the simulations via a relational database launched at runtime, and synthetic contact networks, which

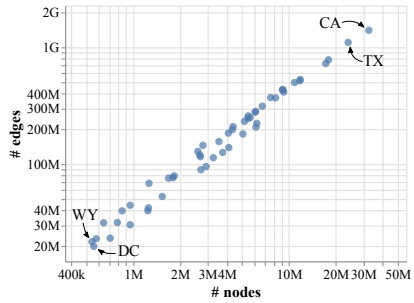


Fig. 3: A scatter plot of the number of nodes and edges in the social contact network for each state in the synthetic population.

are partitioned pre-simulation and pre-loaded into the memory of allocated compute nodes in order to support scalability. The simulation also keeps track of the health state of each individual at each tick (representing one day). We refer the readers to Appendix XII in [1] for further details about the simulation platform.

B. Synthetic populations and social contact networks

EpiHiper simulations depend upon *detailed synthetic populations* and *contact networks* to support accurate and realistic simulations. The synthetic populations are prepared for each of the 50 US states, as well as Washington DC. Figure 3 shows a scatter plot of the node and edge counts by US state.

For each population, the synthetic population data contains the *traits* of each synthetic person. Whereas particular sets of traits may vary across simulations, typical choices for the US include household ID, age and age group, gender, county code, and the latitude and longitude of home locations.

The agent-based models use dynamic contact networks to encode interactions between people during simulations. The initial dynamic contact network in EpiHiper is generated statically. However, during the course of the simulation, each edge in the contact network can be turned on and off dynamically as required in response to, for example, social distancing interventions. Each edge in the contact network includes the identifiers of the two people in contact, and is annotated by the duration of the interaction and the context in which the people meet (home, work, shopping, other, school, college, and religion).

Due to the large size of the contact networks, the network is partitioned between different MPI processes at the beginning of the simulation run. The overall objective is to split the contact network so each partition contains approximately the same number of edges, while at the same time ensuring that all incoming edges of any given node are in the same partition.

Each individual $p \in P$ in the synthetic population is assigned a week-long activity sequence $\alpha(p) = (ai, p)$ where each activity (ai, p) has a start time, a duration, and an activity type from A, where A is the set of activity types. Data sources used for this step include the *National Household Travel Survey* (NHTS) [38], the *American Time Use Survey* (ATUS) [37] and the *Medical Treatment Utilization Schedule* (MTUS) [36], and these are fused to form consistent, week-long activity sequences. We write $\alpha : P \rightarrow A$ for the mapping

	Bridges 2	Rivanna
# Total nodes	488	115
# Allocated nodes	50	50
# CPUs/node	2	2
# Cores/CPU	64	20
RAM per node	256GB (DDR4)	384GB (DDR4)
CPU	AMD EPYC 7742	Intel Xeon Gold 6148
Network	Mellanox ConnectX-6	Mellanox ConnectX-5
OS	CentOS Linux 8	CentOS Linux 7
Filesystem	Lustre	Lustre

TABLE I: Configuration of Bridges 2 and Rivanna.

assigned to each person. For this construction, we use Fitted Values Matching (FVM) for adults [26], and Classification And Regression Tree (CART) for children (see, e.g., [5]).

C. *Bridges 2 and Rivanna: A multi-HPC cluster setup*

Due to the compute-intensive nature of the simulations, the cluster at the author's home university was not sufficient to run the workflows in 'real-time'. For this work, we assume a time limit of 2-3 days as the real-time limit to run the simulations, as further time is needed for human evaluation of the results in order to support weekly delivery schedules for our stakeholders.

Thus, the methodology presented in this paper makes use of two computing clusters, Bridges 2 and Rivanna. Rivanna refers to the computing cluster available at the University of Virginia, and is a modest-sized cluster relative to the significantly more powerful Bridges 2 cluster which is housed at the Pittsburgh Supercomputing Center (PSC). Both clusters use SLURM[1] for scheduling. It should be noted that we were only allocated a subset of Bridges 2's nodes on its clusters due to availability issues.

Job submission system. To be able to utilize the SLURM-driven heterogeneous multi-cluster setup, we implemented a custom job submission system for our workflow. The system assumes that simulation jobs are organized by the regions they will be simulating. For each of the regions, a set of simulation configurations (cells) define the overall parameters of the simulation. However, since the simulations are stochastic in nature, each simulation configuration needs to be executed multiple times to obtain robust statistical measures. Each of these simulations is called a replicate. From the perspective of the submission system, each such replicate is a task, and the overall objective given a set of tasks is to be able to execute all of them in a minimum amount of time given a set of compute nodes.

At a high level, the system simulates one region at a time. Before the simulations for a region are started, a PostgreSQL database instance serving the data for the region is started. The database instances are assigned a full compute node to run. It is assumed that simulations for a given region take roughly the same amount of time and require the same amount of CPU and memory. Since on the Bridges 2 cluster the simulations are memory-bound, we use a simple heuristic to estimate the amount of memory required to run the simulations. Every simulation is assigned at least $k \times E$ memory where E is the size of the contact network for the region (in GB) and k is a heuristic constant. On Bridges 2 we use $k = 8$. When no more simulations for a given region are remaining, the database server for the next region is started, followed by the simulations for this next region. The regions are ordered by the resource requirements (largest regions first).

Phase	CPU usage (core hours)	Memory usage (GB hours)
Simulation	243,200	486,400
Postprocessing	33,024	66,048
Data compression	298	596
Total	276,522	553,044

TABLE II: Total compute time (in core hours) and memory usage (in GB hours) used to complete a US national run (50 states and Washington DC) with a 15-cell and 15-replicate setup on Bridges 2.

The job submission system itself is run as a SLURM job and is written in Python. The job submission system monitors the SLURM job queue to start new simulations and database jobs when necessary, and to stop database jobs when no more simulations need the resources that it is serving.

Table II shows the total compute time (in core hours) and memory usage (in GB hours) used to complete a US national run (50 states and Washington DC) with a 15-cell and 15-replicate setup on Bridges 2.

To transfer data between the two clusters we use the Globus service[2].

V. Experiment Settings and Design

For the experiments, we use an agent-based simulation model, EpiHiper (see Section IV-A). The simulation's input parameters specify the population demographics and contact network, COVID-19 disease model, initial configuration S_0, non-pharmaceutical interventions (NPIs), and vaccination schedule. The simulation output is a dendrogram: a directed graph that tells us who infects whom and on what day. From the output data, we can compute many epidemiological measures, including daily new infections, cumulative infections, prevalence in each state/county, hospitalizations, and deaths, as well as many other outcome measures.

A. *Simulation parameterization*

Our simulation experiment uses synthetic populations and contact networks for the 50 US states and DC. The initial conditions are calibrated to the conditions in each state as of February 7, 2021. Every simulation is run for 300 days. To address the stochasticity in the simulation outcomes, 15 replicates of each simulation are run, and distributions of the outcome measures are computed. All figures presented in

[1]https://slurm.schedmd.com/

[2]https://www.globus.org/

Section VI are based on data from 15 replicates. The curves show an uncertainty of one standard deviation above and below the mean.

Disease model. The disease model is the *best guess version* of "COVID-19 Pandemic Planning Scenarios" prepared by the US Centers for Disease Control and Prevention (CDC) SARS-CoV-2 Modeling Team [10], and has been used by multiple researchers in their papers. It is a Susceptible-Exposed-Infectious-Recovered (SEIR) model where state transitions follow the parameters as defined in the document. The disease states and transition paths are shown in Figure 9 in [1]. Individuals of different age groups have different infectivity and susceptibility; dwell time distributions and state transition probability distributions are stratified by the following age groups: preschool (0–4 years), students (5–17), adults (18–49), older adults (50–64) and seniors (65+). Furthermore, individuals who are vaccinated have different disease parameter values than those who are not vaccinated. Detailed parameterization for unvaccinated individuals is summarized in Appendix XI (see [1]).

To model the fast spread of new variants in the population, we increase the transmissibility in the disease model over time.

Initializations. The simulations are initialized at the county level by age group using the detailed data of confirmed cases from [35]. The initialization specifies the health state of each individual. Based on county-level cumulative confirmed cases through January 21, 2021, we derive the number of prior infections in each county by scaling the cumulative number by a case ascertainment ratio of 3 (i.e., only one third of all infections are reported), then computing the number of prior infections in each age group of this county using the age distribution in cases. We randomly choose individuals in each age group in each county and set their health states to `recovered` to reflect that they have already been infected. Based on county-level daily confirmed cases from January 22 to February 7, 2021, we derive the number of individuals who are infected each day by the same scaling, and seed the simulation by setting randomly chosen individuals to *exposed* by day in each age group of each county.

Non-pharmaceutical interventions. We consider four NPIs: (i) *Infectivity reduction (IR)*. A fraction (75%) of the population chooses to practice preventive behavior, e.g., mask wearing and hand washing, which reduces their infectivity by 60%, if they are infected. (ii) *Generic social distancing (GSD)*. A fraction (15%) of the population chooses to reduce non-essential (shopping, religion, and other) activities. (iii) *Virtual learning (VL)*. A fraction (25%) of K-12 students choose virtual learning. (iv) *Voluntary home isolation of symptomatic cases (VHI)*. With a probability of 75%, a symptomatic person chooses to stay home for 14 days, reducing the weights on household contacts by 50%. For this person, all outside contacts are disabled, and at-home contacts are temporarily reduced by 50% during these 14 days.

B. Vaccination

Vaccine efficacy. Overall vaccine efficacy is characterized by three numbers: (i) e_I, efficacy against infection; (ii) e_D, efficacy against severe illness (requiring hospitalization or leading to death) given infection; and (iii) e_T, efficacy against onward transmission given infection. Efficacy differs for different vaccines. In this study we assume a hypothetical vaccine similar to Pfizer-BioNTech and Moderna which are the main vaccines administered in the US. We ignore the partial efficacy after the first dose takes effect, and assume that $e_I = 95\%$, $e_D = 50\%$, and $e_T = 100\%$ starting only 21 days after full vaccination.

Vaccine hesitancy. Not everyone is willing to take COVID-19 vaccines. People may hesitate to get vaccinated for a variety of reasons, including medical, social, ideological, or religious concerns. The acceptance rate differs in different populations. We take the recent Facebook survey results [13] on vaccine hesitancy in each state, and use the acceptance rate as the upper bound of the total number of people who will be fully vaccinated in that state. This upper bound varies from below 50% (e.g. in Mississippi) to above 90% (e.g. in New Hampshire) – see Figure 11 in [1] for more details.

Vaccine demand schedule. We assume that vaccination in the US is constrained by the demand (acceptance) but not by the supply. That is, vaccines are available as long as there are people willing to get vaccinated. Regarding vaccination rate, we assume two different schedules: (i) *accelerated* vaccination and (ii) *accelerated-decelerated* demand in vaccination. The former schedule has a state-specific constant acceleration each week such that all people willing to be vaccinated in the state are vaccinated in 20 weeks (by the end of May 2021). The latter schedule has a smaller constant acceleration each week in the first 20 weeks (by the end of May 2021) but starts to decelerate in the next 20 weeks (from June to October 2021) due to gradual saturation towards the acceptance upper bound. Note that the weekly vaccination rates, counts, and percentage of population depend on the different vaccine acceptance rate in each state. We have also considered a scenario where the vaccine acceptance rate is increased by 10% in every state. This can be achieved by improving the convenience of getting vaccinated, transparency of information on vaccination outcomes including side effects, and social contagion of vaccination behavior via various social networks.

Vaccination prioritization. We assume that vaccines are only given to people who are at least 18 years old. Among those people, we consider the following prioritization strategies.

- *No priority*. Everyone 18+ years old is vaccinated with a random ordering, given that they are willing to be vaccinated. This is our baseline strategy.
- *Age-based priority*. This strategy prioritizes those who are at least 50 years old (older adults and seniors). In all vaccines administered each week, most (80%) are allocated to the prioritized groups, but allow some (20%) to be given to adults (18–49).

C. Experimental design

The design consists of 3 factors: (i) 2 vaccine acceptance levels: survey-based acceptance rate and improved acceptance rate (improved by 10%); (ii) 2 vaccination demand schedules: accelerated (accel) and accelerated-decelerated (accel-decel); (iii) 2 prioritization schemes: no priority and age-based prioritization. Combining the three factors, we have 8 cells, plus a cell (*no-vax*) where the NPIs are in place, but the vaccinations are not applied.

Calibration. We calibrate the disease model using the *no-vax* cell for each state, where we fix the initializations and NPIs, and all parameters in the disease model except for *transmissibility*, a parameter representing the generic disease infectiousness. For each state, we search for the transmissibility by targeting the effective reproduction number, $R_{\text{effective}}$, estimated at the beginning of February 2021 using state level confirmed case counts up to that time. We use the binary search method, where in each iteration we update transmissibility and run simulation with the *no-vax* settings, then estimate $R_{\text{effective}}$ from the simulation generated case counts and compare it with the target value. Our search converges when the simulation-based $R_{\text{effective}}$ matches the ground truth-based $R_{\text{effective}}$. The same disease model is then used for all of the simulations in the remaining 8 cells. Each simulation runs for 300 days and for 15 replicates.

VI. Results and Analysis

Effectiveness of vaccinations differs in different states. Figure 4 shows that, in terms of averted cases/deaths, vaccination is more effective in some states than in others. Note that, for comparison between states, the averted infections and deaths are normalized by population size, so the numbers are per 100K population. In the figure, the dashed lines are the national average of averted infections (1364) and averted deaths (9), normalized by the national population.

It is not obvious that averted infections/deaths are higher in states with higher vaccine acceptance. For example, while Hawaii and Rhode Island rank top in both vaccine acceptance and infection/death aversion, California and Massachusetts rank top in vaccine acceptance but their infection/death aversions are below the national average. On the other hand, Wyoming, Mississippi, and Oklahoma have the lowest vaccine acceptance rates, but their infection/death aversions are above the national average. In general, states with large populations, like California and Texas, have a smaller number of averted infections/deaths. There is a small positive correlation between averted infections/deaths and vaccine acceptance, but a significant negative correlation between averted infections/deaths and population size, as shown in Table III.

Vaccine hesitancy reduces the effectiveness of vaccinations. At the national level, the accelerated-decelerated vaccination demand due to vaccine hesitancy leads to more infections/deaths (fewer averted infections/deaths), compared to the accelerated vaccination schedule. This is shown in Figure 5. With an accelerated vaccination schedule, 2042 infections and 12 deaths can be averted per 100K people. At the national level, this results in a total reduction of 6.7M infections and 39.4K deaths. With an accelerated-decelerated vaccination demand schedule, the averted infections and deaths decrease to 1364 and 8.6, respectively, per 100K. At the national level it reduces 4.5M infections and 28.2K deaths in total.

vaccination	acceptance vs. aversion of		population size vs. aversion of	
	infection	death	infection	death
no priority				
accel	0.050	0.042	-0.475	-0.445
accel-decel	0.078	0.079	-0.471	-0.439
age-based				
accel	0.040	0.086	-0.482	-0.445
accel-decel	0.088	0.085	-0.473	-0.448

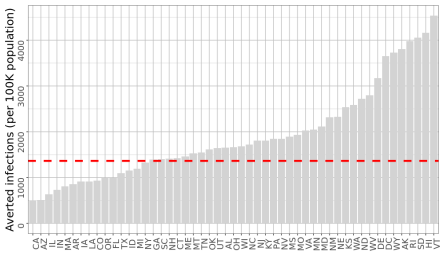
TABLE III: Correlation between averted infections / deaths and vaccine acceptance / population size, under different vaccination prioritizations and schedules.

These numbers highlight the health and human costs of a slower vaccination schedule due to saturation in demand for vaccination and vaccine hesitancy. Although the same number of people are vaccinated at the end of both vaccination schedules, because vaccines are administered faster in one schedule than the other, the vaccines protect more people in the accelerated scenario through indirect protections, i.e., more people avoid getting infected before getting vaccinated.

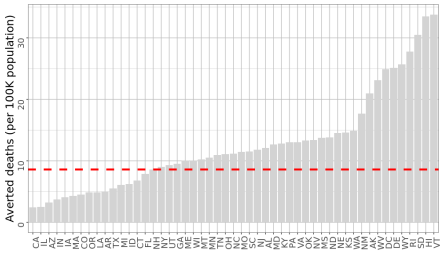
Infections and deaths can be further reduced by improving vaccine acceptance. Suppose we can increase the vaccine acceptance rate by 10% in each state. Figure 6 shows that this can improve aversion of infections and deaths over the effect of vaccination at the current acceptance level. With an accelerated-decelerated vaccination schedule, 10% more vaccine acceptance can increase infection aversion from 1364 to 1445 and death aversion from 8.6 to 9.1 per 100K of the national population, corresponding to a total aversion of 4.7M infections and 29.9K deaths. Although a smaller effect compared to what an accelerated vaccination schedule can achieve, the improvement is still significant.

VII. Discussion

To get a closer look at the impact of slower vaccination rates and the benefit of improved vaccine acceptance, we compare the infection aversion in three scenarios, (i) accelerated vaccination, (ii) accelerated-decelerated vaccination, and (iii) accelerated-decelerated vaccination with 10% more vaccine acceptance, for each state and compare the results across states (see Figure 7 in [1] for the detailed plot). We find that while the observations in Sections VI and VI hold in all states, the magnitude of the impact is heterogeneous between states. For example, in the northeast states, the benefit of improved vaccine acceptance is only marginal, whereas in Louisiana, Mississippi, and West Virginia, it significantly helps in averting infections. This is because marginal improvements from higher vaccine acceptance in cases and deaths depend on the current level of acceptance. In the aforementioned states, the vaccine acceptance and uptake levels are very low

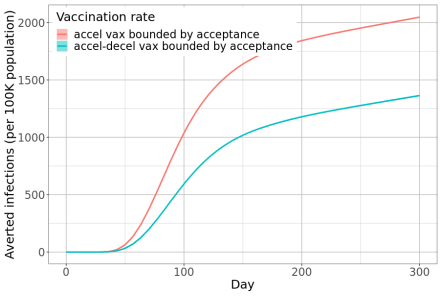


(a) Averted infections

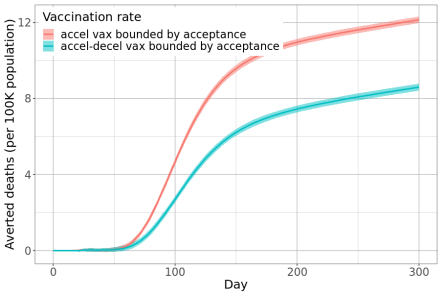


(b) Averted deaths

Fig. 4: Total averted infections (a) and averted deaths (b) can be very different between states. The bars show the average averted totals in each state, normalized to count per 100K population. The red horizontal lines show the average averted totals in the whole US.

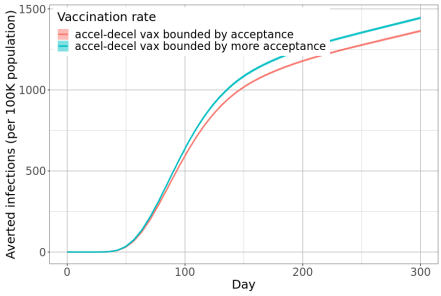


(a) Averted infections

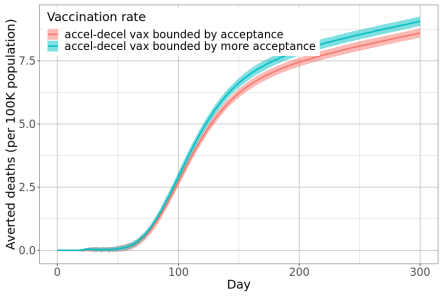


(b) Averted deaths

Fig. 5: Due to vaccine saturation, vaccination rates decrease after the initial acceleration. This leads to fewer averted infections (a) and fewer averted deaths (b) compared with a constant acceleration of vaccination rate.



(a) Averted infections



(b) Averted deaths

Fig. 6: If we can increase vaccine acceptance by 10%, we can reduce cases further in terms of (a) infections and (b) deaths.

compared to the other states. Therefore, an additional 10% vaccine acceptance would make a much bigger difference than in a state where the current acceptance and uptake levels are already high.

We can investigate such effects at sub-state levels. For example if we zoom in on Virginia and compare the impact between different districts, we find that while the benefit of improved vaccine acceptance is significant in the northern and eastern districts of Virginia, it is very limited in southwest Virginia (see Figure 7 in [1] for the detailed plot).

To further investigate the heterogeneity of the infection/death aversion between states, we compare the averted infections in each state (with accelerated-decelerated vaccination rates) with vaccine acceptance as well as demographic attributes, such as population size, average age, and average household size (see Figure 8 in [1] for the detailed scatter plot). We find that the states have heterogeneous attributes, and that the averted infections seem to be more correlated to population size and average household size than to vaccine acceptance or average age.

VIII. Summary and Conclusions

This paper studies the role of vaccine acceptance in controlling the spread of COVID-19 in the US. Carrying out the massive simulations, necessary for this study, required high performance computing resources and addressing big-data challenges. We showed how complex epidemiological workflows can be carried out by simultaneously using two large HPC systems; the need for this arose due to the machine configurations, access priority, and fast turnaround times needed by federal and state agencies to support their COVID-19 response efforts. We consider our vaccine study workflow to be an exemplar case study demonstrating how scientific productivity can be improved by developing and managing better scientific workflows, which has recently been recognized as an important issue in the HPC community [20]. Our results show that vaccine acceptance has significant impacts on the overall outcome of vaccination, and that this impact is spatially and temporally diverse. Certain states are likely to do better even with lower levels of acceptance than other states. The results also quantify the marginal utility one can get by increasing vaccine acceptance levels.

Acknowledgements The authors would like to thank members of the Biocomplexity COVID-19 Response Team, Network Systems Science and Advanced Computing (NSSAC) Division, UVA Research computing, Pittsburgh Supercomputing Center, John Towns and our partners at CDC, VDH, NSF. This work was partially supported by the National Institutes of Health (NIH) Grant R01GM109718, VDH Grant PV-BII VDH COVID-19 Modeling Program VDH-21-501-0135, NSF Grant No.: OAC-1916805, NSF Expeditions in Computing Grant CCF-1918656, CCF-1917819, NSF RAPID CNS-2028004, NSF RAPID OAC-2027541, US Centers for Disease Control and Prevention 75D30119C05935, DTRA subcontract/ARA S-D00189-15-TO-01-UVA, NSF XSEDE TG-BIO210084. This work used resources, services, and support from the COVID-19 HPC Consortium (https://covid19- hpc-consortium.org/).

REFERENCES

[1] AI-Driven Agent-Based Models to Study the Role of Vaccine Hesitancy in Controlling COVID-19 in the US: Extended Version. https://bit.ly/325qslF.

[2] P. Bhattacharya, J. Chen, S. Hoops, D. Machi, B. Lewis, S. Venkatramanan, M. L. Wilson, B. Klahn, A. Adiga, B. Hurt, J. Outten, A. Adiga, A. Warren, H. Baek, P. Porebski, A. Marathe, D. Xie, S. Swarup, A. Vullikanti, H. Mortveit, S. Eubank, C. L. Barrett, and M. Marathe. Data-Driven Scalable Pipeline using National Agent-Based Models for Real-time Pandemic Response and Decision Support. In *Finalist for the 2021 ACM Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research*, 2021.

[3] K. R. Bisset, J. Chen, X. Feng, V. A. Kumar, and M. V. Marathe. Epifast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *Proceedings of the 23rd international conference on Supercomputing*, pages 430–439, 2009.

[4] R. K. Borchering, C. Viboud, E. Howerton, C. P. Smith, S. Truelove, M. C. Runge, N. G. Reich, L. Contamin, J. Levander, J. Salerno, et al. Modeling of future covid-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios—united states, april–september 2021. 2021.

[5] L. Breiman. *Classification and regression trees*. Wadsworth statistics/probability series. Wadsworth International Group, 1984.

[6] K. M. Bubar, K. Reinholt, S. M. Kissler, M. Lipsitch, S. Cobey, Y. H. Grad, and D. B. Larremore. Model-informed covid-19 vaccine prioritization strategies by age and serostatus. *Science*, 371(6532):916–921, 2021.

[7] J. H. Buckner, G. Chowell, and M. R. Springborn. Dynamic prioritization of covid-19 vaccines when social distancing is limited for essential workers. *Proceedings of the National Academy of Sciences*, 118(16), 2021.

[8] U. CDC. Trends in Number of COVID-19 Vaccinations in the US. https://covid.cdc.gov/covid-data-tracker/#vaccination-trends.

[9] Cdc/atsdr social vulnerability index. https://www.atsdr.cdc.gov/placeandhealth/svi/index.html.

[10] Centers for Disease Control and Prevention. Covid-19 pandemic planning scenarios. https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html, 2020. [Online, accessed May 10, 2021].

[11] J. Chen, S. Hoops, A. Marathe, H. Mortveit, B. Lewis, S. Venkatramanan, A. Haddadan, P. Bhattacharya, A. Adiga, A. Vullikanti, et al. Prioritizing allocation of covid-19 vaccines based on social contacts increases vaccination effectiveness. *medRxiv*, 2021.

[12] J. Chen, A. Vullikanti, S. Hoops, H. Mortveit, B. Lewis, S. Venkatramanan, W. You, S. Eubank, M. Marathe, C. Barrett, et al. Medical costs of keeping the us economy open during covid-19. *Scientific reports*, 10(1):1–10, 2020.

[13] COVIDcast: Vaccine Acceptance Summary. https://delphi.cmu.edu/covidcast/indicator/?sensor=fb-survey-smoothed_covid_vaccinated_or_accept&date=20210521.

[14] A. A. Dror, N. Eisenbach, S. Taiber, N. G. Morozov, M. Mizrachi, A. Zigron, S. Srouji, and E. Sela. Vaccine hesitancy: the next challenge in the fight against covid-19. *European journal of epidemiology*, 35(8):775–779, 2020.

[15] S. Eubank, H. Guclu, V. S. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:180–184, 2004.

[16] M. C. Fitzpatrick and A. P. Galvani. Optimizing age-specific vaccination. *Science*, 371(6532):890–891, 2021.

[17] I. Foster, D. Parkes, and S. Zheng. The rise of ai-driven simulators: Building a new crystal ball. *arXiv preprint arXiv:2012.06049*, 2020.

[18] J. R. Goldstein, T. Cassidy, and K. W. Wachter. Vaccinating the oldest against covid-19 saves both the most lives and most years of life. *Proceedings of the National Academy of Sciences*, 118(11), 2021.

[19] M. E. Halloran, N. M. Ferguson, S. Eubank, I. M. Longini, D. A. T. Cummings, B. Lewis, S. Xu, C. Fraser, A. Vullikanti, T. C. Germann, D. Wagener, R. Beckman, K. Kadau, C. Barrett, C. A. Macken, D. S. Burke, and P. Cooley. Modeling targeted layered containment of an influenza pandemic in the united states. *Proceedings of the National Academy of Sciences*, 105(12):4639–4644, 2008.

[20] M. A. Heroux, L. McInnes, D. Bernholdt, A. Dubey, E. Gonsiorowski, O. Marques, J. D. Moulton, B. Norris, E. Raybourn, S. Balay, R. A. Bartlett, L. Childers, T. Gamblin, P. Grubel, R. Gupta, R. Hartman-Baker, J. Hill, S. Hudson, C. Junghans, A. Klinvex, R. Milewicz, M. Miller, H. Ah Nam, J. O Neal, K. Riley, B. Sims, J. Schuler, B. F. Smith, L. Vernon, G. R. Watson, J. Willenbring, and P. Wolfenbarger. Advancing Scientific Productivity through Better Scientific Software: Developer Productivity and Software Sustainability Report.

[21] A. Hogan, P. Winskill, O. Watson, P. Walker, C. Whittaker, M. Baguelin, D. Haw, A. Lochen, K. Gaythorpe, K. Ainslie, et al. Report 33: Modelling the allocation and impact of a covid-19 vaccine. 2020.

[22] P. C. Jentsch, M. Anand, and C. T. Bauch. Prioritising covid-19 vaccination in changing social and epidemiological landscapes: a mathematical modelling study. *The Lancet Infectious Diseases*, 2021.

[23] R. M. Kaplan and A. Milstein. Influence of a covid-19 vaccine?s effectiveness and safety profile on vaccination acceptance. *Proceedings of the National Academy of Sciences*, 118(10), 2021.

[24] J. C. Lemaitre, D. Pasetto, M. Zanon, E. Bertuzzo, L. Mari, S. Miccoli, R. Casagrandi, M. Gatto, and A. Rinaldo. Optimizing the spatio-temporal allocation of covid-19 vaccines: Italy as a case study. *medRxiv*, 2021.

[25] E. T. Lofgren, M. E. Halloran, C. M. Rivers, J. M. Drake, T. C. Porco, B. Lewis, W. Yang, A. Vespignani, J. Shaman, J. N. Eisenberg, et al. Opinion: Mathematical models: A key tool for outbreak response. *Proceedings of the National Academy of Sciences*, 111(51):18095–18096, 2014.

[26] K. Lum, Y. Chungbaek, S. Eubank, and M. Marathe. A two-stage, fitted values approach to activity matching. *International Journal of Transportation*, 4:41–56, 2016.

[27] D. Machi, P. Bhattacharya, S. Hoops, J. Chen, H. Mortveit, S. Venkatramanan, B. Lewis, M. Wilson, A. Fadikar, T. Maiden, et al. Scalable epidemiological workflows to support covid-19 planning and response. *medRxiv*, 2021.

[28] L. Matrajt and I. M. Longini Jr. Optimizing vaccine allocation at different points in time during an epidemic. *PloS one*, 5(11):e13767, 2010.

[29] J. Medlock and A. P. Galvani. Optimizing influenza vaccine distribution. *Science*, 325(5948):1705–1708, 2009.

[30] S. M. Moghadas, T. N. Vilches, K. Zhang, S. Nourbakhsh, P. Sah, M. C. Fitzpatrick, and A. P. Galvani. Evaluation of covid-19 vaccination strategies with a delayed second dose. *PLoS Biology*, 19(4):e3001211, 2021.

[31] S. Moore, E. M. Hill, M. J. Tildesley, L. Dyson, and M. J. Keeling. Vaccination and non-pharmaceutical interventions for covid-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 2021.

[32] A. Perrault, F. Fang, A. Sinha, and M. Tambe. Ai for social impact: Learning and planning in the data-to-deployment pipeline. *arXiv preprint arXiv:2001.00088*, 2019.

[33] C. M. Rivers, E. T. Lofgren, M. Marathe, S. Eubank, and B. L. Lewis. Modeling the impact of interventions on an epidemic of ebola in sierra leone and liberia. *PLoS currents*, 6, 2014.

[34] M. Sallam. Covid-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. *Vaccines*, 9(2):160, 2021.

[35] The New York Times. Coronavirus (covid-19) data in the United States. https://github.com/nytimes/covid-19-data, last accessed on February 7, 2021, 2020.

[36] The University of Oxford. The Multinational Time Use Study (MTUS). Last accessed: February 2020.

[37] United States Department of Labor, Bureau of Labor Statistics. The American Time Use Survey (ATUS). Last accessed: February 2020.

[38] U.S. Department of Transportation, Federal Highway Administration. The National Household Travel Survey (NHTS). Last accessed: February 2020.

[39] S. Venkatramanan, J. Chen, A. Fadikar, S. Gupta, D. Higdon, B. Lewis, M. Marathe, H. Mortveit, and A. Vullikanti. Optimizing spatial allocation of seasonal influenza vaccine under temporal constraints. *PLoS computational biology*, 15(9):e1007111, 2019.

[40] E. Wrigley-Field, M. V. Kiang, A. R. Riley, M. Barbieri, Y.-H. Chen, K. A. Duchowny, E. C. Matthay, D. Van Riper, K. Jegathesan, K. Bibbins-Domingo, et al. Geographically-targeted covid-19 vaccination is more equitable than age-based thresholds alone. *medRxiv*, 2021.