BI Technical Report 2024-33

September 12th, 2024

Title: A Digital Twin of an Evolving Pandemic for Genomic Epidemiology

Authors: Biocomplexity Institute, Team

Contact: Jiangzhuo Chen, Madhav Marathe
Email: chenj@virginia.edu,marathe@virginia.edu

Status: This technical report is delivered together with a synthetic pandemic data set to CDC.

Acknowledgements: This work has been partially supported by the Centers for Disease Control and Prevention (CDC) through Pathogen Genomics Centers of Excellence network (PGCoE) grant 6NU50CK000555-03-01, National Science Foundation (NSF) grants CCF-1918656 (Expeditions), OAC-1916805 (CINES), and DMS-2327710 (IHBEM), the Defense Threat Reduction Agency (DTRA) contract HDTRA120F0017.

A Digital Twin of an Evolving Pandemic for Genomic Epidemiology

Biocomplexity Institute, University of Virginia

1 Introduction

A synthetic pandemic data set is an extension of a base *digital twin*. The base digital twin of a particular population, e.g., Virginia, is a synthetic representation of the population. It consists of synthetic individuals with demographic, geographic, and socioeconomic attributes, and their activities and mobility with spatio-temporal labels. On top of it, we can add additional layers of synthetic data to enable the study of various problems and applications. For example, we have already added a *social contact network* layer to represent the physical proximity of individuals when they visit and stay at the same location at the same time. This contact network data allows us to model infectious disease dynamics in the population such as influenza, Ebola, and COVID-19. Such additional layers are attached to a particular version of the base digital twin and become part of the digital twin.

In this work, we add another data layer to the digital twin using an agent-based epidemiological model, to represent an infectious disease outbreak. We use COVID-19 as a specific example of an infectious disease here; extensions to other airborne viral diseases can be undertaken by extending these basic techniques. We provide several sets of pandemic data to allow the study of various problems such as bio-surveillance, optimal resource allocation, intervention assessment, and cascade reconstruction to name a few. All data sets are based on and are extensions of our v1.9.0 digital twin of Virginia, which consists of a synthetic population of Virginia with demographic, geographic, and socio-economic attributes of each individual and each household, and a synthetic contact network with labeled nodes representing synthetic individuals and labeled edges representing physical proximity between individuals enabling disease spread. The synthetic outbreak data represents instances of realistic epidemic outbreak over the synthetic social contact network based on realistic COVID-19 models of within-host disease progression and between-host disease transmission.

In these synthetic pandemic data sets, we model different scenarios corresponding to different stages of the COVID-19 pandemic. Each data set highlights specific features. The first set models the period when the Delta variant emerged and co-circulated with the Alpha variant, and highlights between-variant competition and vaccine hesitancy; the second set models the period when a large percentage of the population had been infected and/or vaccinated and had obtained some level of immunity, and features waning of the natural and vaccinal immunity in the population; the third set models the period when many new variants, including Omicron, emerged and many people had been reinfected (some more than once) and highlights immune escape and the endemicity of the COVID-19 pandemic. These scenarios are motivated by the scenario model hub (SMH) efforts [1, 3, 4, 2] in which we have been participating since 2021 [12].

2 Base Data Set: Digital Twin of Virginia

Our group has been developing methods for creating synthetic population data sets for over 25 years [25, 5, 19, 15]. In our efforts to support the federal government and the Virginia Health Department in their response to the COVID-19 pandemic, we created a synthetic social contact network of the state of Virginia, as well as other parts of the United States (US). In the network, nodes represent individuals, and edges capture physical proximity. Using these networks, we created epidemic simulations based on real-world outbreak information on vaccinations, non-pharmaceutical interventions, and other relevant data. The output of one simulation is a set of synthetic individuals who are infected over the course of the disease transmission. The

synthetic data (also referred to as a *digital twin* in recent literature) provides a realistic account of how the disease spreads through the population in time and space.

A synthetic agent (e.g., person) is assigned states and interactions that make it statistically consistent with members of the (real) population without necessarily matching the characteristics of any specific (real) person. A synthetic population represents a set of synthetic agents (e.g., people) that share common geographic or social characteristics (e.g., people in a rural or urban region, individuals from a specific state).

These populations and networks are formed by collecting a large and diverse set of publicly and/or commercially available data sets. These data sets include census, land use, mobility, activity, behavioral and transportation surveys and building maps. The data sets have been integrated in a first principles manner to construct these synthetic populations. They have been used for highly accurate, national level, agent-based modeling tasks [8].

The synthetic data is formed by starting with the empirical distribution of particular attributes within an area (for example, the number of people per household, age, or income), then iteratively forming a population that matches those distributions while also preserving observed associations between those attributes. We apply a similar process to mobility data. This process means that there is no direct correspondence between any real person and any single synthetic agent within our synthetic population.

Our methodology ensures that privacy is maintained. We apply our synthesizing process to data that is already public, and, in the case of the census, has already had privacy-enhancing methods applied to it [18]. In the case of the 2020 census, the data that was released to the public had differential privacy applied to it [31], and our use corresponds to post-processing, which does not have any additional impact on privacy [14]. Furthermore, when we do use sources that have not been subjected to differential privacy processing, we rely on data sources that are already public. Any data an attacker might attempt to derive from the synthetic data could be much more easily obtained by looking at the sources we used to construct the data in the first place. Therefore, our digital twin and synthetic pandemic data sets preserve privacy.

The base synthetic population data of Virginia which we provide consists of synthetic individuals, each of whom is assigned an age, a gender, a household and its home location, visited locations, and the activities performed at those locations. The synthetic network data of Virginia we provide consists of a person-location graph (also called an activity-location assignment graph) and a person-person contact network that is derived from the person-location graph based on which individuals visit the same location at the same time.

2.1 Methodology to Generate Synthetic Population and Network of Virginia

A *synthetic population* of a region may be regarded as a digital twin of the real population of that region. In this section, we provide a compact summary of the models and methodologies behind constructing synthetic populations and contact networks. See [23] for additional details. Our work builds on earlier techniques using a first principles approach for constructing synthetic populations [16, 17, 6].

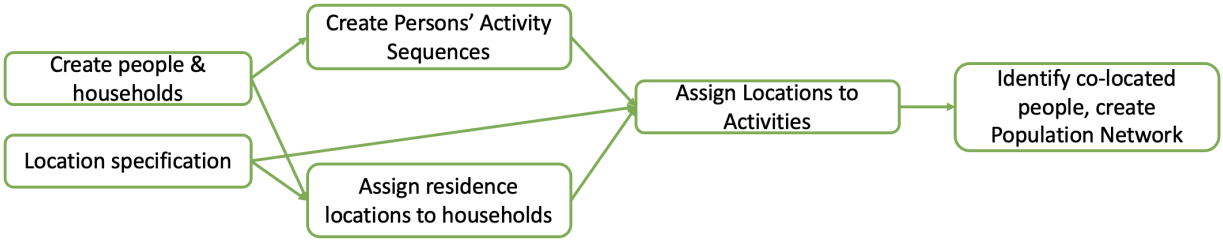


Figure 1: High-level sequence of models and steps used for constructing synthetic populations.

To construct a population for a *geographic region* R (e.g., Virginia), we first choose a collection of *person attributes* from a set $\mathcal{D}$ (e.g., age and gender) and a set $\mathcal{T}_A$ of *activity types* (e.g., `Home`, `Work`, `Shopping`, `Other`, `School`, `College`, and `Religion`). The precise choices of $\mathcal{D}$ and $\mathcal{T}_A$ are guided by the particular

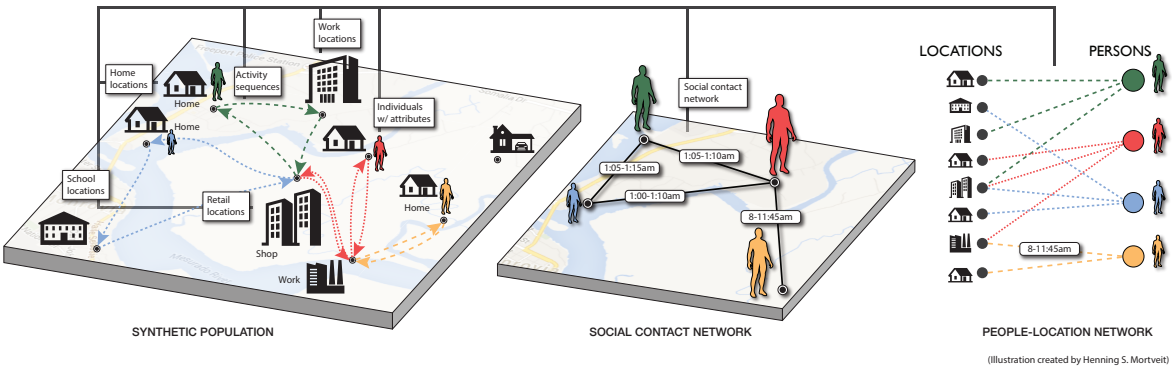


Figure 2: A high-level illustration showing the manner in which data is integrated in the modeling and construction of synthetic populations. The members of the populations will be equipped with a range of demographic attributes (details will depend on available data and application purpose), will have an associated contact network (denoted by G_P) as shown in the middle, and a person-location graph (denoted by G_{PL}) as shown on the right.

scenarios or analyses the population will serve. At a high level (see Figure 2), we: (*i*) construct virtual people and places, (*ii*) assign activity sequences to people, (*iii*) assign each activity a location and time of visit, and from this, we derive (*iv*) a contact network using co-occupancy and a contact model to infer edges. A high-level workflow for this process is illustrated in Figure 1 where the contact network is illustrated in the middle. The construction factors into a detailed sequence of steps which can be outlined as follows.

Step 1: Person and Household construction. Using *iterative proportional fitting* (IPF) [7, 13], the **base population** model constructs a set of individual persons $\mathcal{P}$ where each person is assigned demographic attributes from $\mathcal{D}$. By design, this ensures that $\mathcal{P}$ matches the statistical distributions of the Public Use Microdata Sample (PUMS) data from the US Census [29], which is one of the input data sets for the model. Additionally, this model partitions $\mathcal{P}$ into a set $\mathcal{H}$ of *households*. Here, the term *household* encompasses the traditional notion of "family" as well as other subsets of individuals residing in the same *dwelling unit* (e.g., dormitories, apartments, army barracks, or prisons).

Step 2: Activity sequence assignment. In this step, each individual $p \in \mathcal{P}$ is assigned a week-long **activity sequence** $\alpha(p) = (a_{i,p})_i$ where each *activity* $a_{i,p}$ has a *start time*, a *duration*, and an *activity type* from $\mathcal{T}_A$. Data sources used as input for this step include the National Household Travel Survey (NHTS) [32], the American Time Use Survey (ATUS) [30] and the Multinational Time Use Study (MTUS) [27]. We write $\alpha\colon \mathcal{P} \longrightarrow \mathcal{A}$ for the mapping assigned to each person. For this construction, we use Fitted Values Matching (FVM) for adults [21], and Classification And Regression Tree (CART) for children (see, e.g.,[9]).

Step 3: Location construction. The **location construction model** generates a set of spatially embedded locations $\mathcal{L}$ partitioned into *residence locations* where households live, and activity locations where people conduct their non-`Home` activities. This construction is highly granular, and is based on several data sources, including the Microsoft US Building data [22], HERE/NAVTEQ [20] for points-of-interest (POIs) and land-use classifications, and the National Center for Education Statistics (NCES) [24] for public schools.

Step 4: Activity location assignment. For each person $p \in \mathcal{P}$, the **activity location assignment** model assigns a location $\ell_i = \ell(a_i)$ to each of their activities a_i. We denote the sequence of locations visited by p as $\lambda_p = (\ell_i)_{i,p}$. The location assignment model uses the American Community Survey (ACS) commute flow data [28] to assign a target county c for each `Work` activity, and a particular location randomly within c based on attractor weights assigned to each location in c. School activity locations are assigned using NCES data, while remaining activities are anchored near home and work locations. The activity location assignment induces the bipartite *people location graph* G_{PL} with vertex sets $V_1 = \mathcal{P}$ and $V_2 = \mathcal{L}$ (the set of locations) and labeled edges all (p, ℓ) for which p visits ℓ. The label includes the activity type, the start time for the visit, and the duration of the visit; for more details, see the right side of Figure 2.

Step 5: Contact network generation. In this step, the **contact network model** uses the people-location graph G_{PL} to first derive the *co-location graph* $G_{\max}$ with vertex set $\mathcal{P}$ and edges all $e = (p, p')$ for people p and p' that are simultaneously present at the same location. Applying sub-location contact modeling at each location, we determine which of the edges of $G_{\max}$ will be retained to form the *contact network* G, which is also referred to as the *person-person contact network* and denoted by G_P (rather than simply G) to make this explicit. In this work, we use a random graph model referred to as the *Min/Max/alpha model* at each location to obtain G_P. Let ℓ be a location and let $N = N_\ell$ denote the maximal number of simultaneous visits to ℓ. We define the function $p_\ell \colon \mathbb{N} \setminus \{0, 1\} \longrightarrow [0, 1]$ as

$$p_\ell(N) = \min\big\{1, \big[\text{Min} + (\text{Max} - \text{Min})(1 - e^{-N/\alpha})\big]/[N - 1]\big\}\,, \tag{1}$$

where Min $<$ Max are non-negative numbers and $\alpha > 0$. Given $p = p_\ell(N)$, one samples from this random graph model in the same manner as for the standard model $G_{n,p}$ by independently applying to each edge e at random the probability p corresponding to the location ℓ where $e \in G_{\max}$ originates. Thus, the parameters Min and Max bound the degree of each vertex locally at ℓ (in expectation) at the beginning of each visit; note, however, that the degree of person p in the resulting graph G is the accumulation of degrees across their trajectory to locations visited while executing their activity sequence. Thus, the choices of Min, Max and α will induce the degree of each vertex in a bottom-up manner; see [23] for full details.

Remark. The Virginia networks G_P feature contacts and edges throughout an entire week. To support this challenge, we extract sub-graphs, e.g., G_1, from G_P to represent the contact network on the particular days.

3 Outbreak Data Sets: Synthetic Pandemic in Virginia

We have highly granular disease models which can be applied to the digital twin of Virginia. A disease model is a probabilistic model of disease progression, specifying the likelihood of transitioning between disease states based on each synthetic agent's attributes. We determine the likelihood of infection given exposure based on a number of factors including social distancing and vaccination status, and can further model whether a specific infection is asymptomatic, or the likelihood that the infection will result in a severe, reported case. Using such a disease model, we generate synthetic outbreak data that reflects how an individual's disease state evolves over time.

3.1 Synthetic Pandemic Data Set 1: An Outbreak with Co-circulating Variants

This data set models the scenario corresponding to the summer of 2021, when a new variant of concern (VOC), Delta, emerged and quickly became dominant in Virginia due to its much higher transmissibility than the circulating variant Alpha. From the end of 2020, vaccines began to be administered to the population, first prioritizing seniors and critical workers, then starting to cover adults. Part of the population had developed immunity from infections by the wild type and the Alpha variant, as well as from early vaccines, including Johnson&Johnson, Pfizer, and Moderna. However, the fraction of the population being infected by COVID-19 was still small, and due to vaccine hesitancy, vaccination coverage was still low by the time the Delta variant emerged and was imported into Virginia. The provided data covers an outbreak of about six months, corresponding to the period between the emergence of the Delta variant and the emergence of the Omicron variant.

What can this data set be used for? In this data set we model two variants: an *existing* variant is seeded on May 5, mainly by the cases already in the population; and a *new* variant is seeded on May 29, by importation. After the emergence of the new variant, the susceptible people of the population are competed for by both variants but the new one is 60% more transmissible than the existing one. The dynamics of the competition varies among different sub-populations and depends on the network structure and the immunity distributions in the contacts of the infectious people of different variants. For example, Figure 3 shows the prevalence of two variants among new infections over time by age group. The new variant becomes dominant

in younger age groups (below 18) about three months after its emergence, but never becomes dominant in older age groups (18+). This data set can be used for studies and analyses related to co-circulating variants. For example, it can be used to test and evaluate surveillance models for detecting an emerging variant and to study budget constrained optimal targeted surveillance problem.

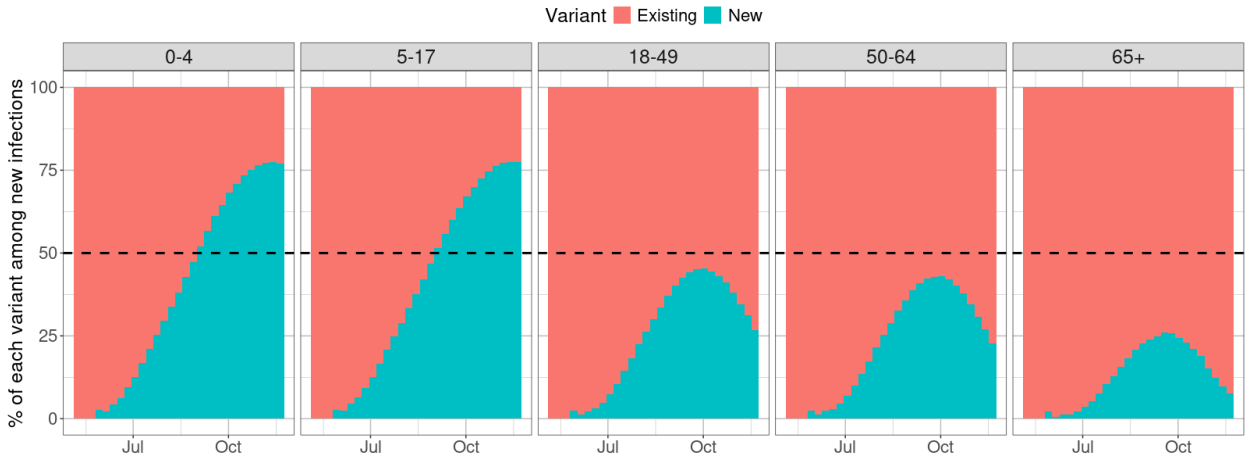


Figure 3: Distribution of new infections between two variants over time. The new variant becomes dominant in preschoolers (0-4) and school-age (5-17) in early September. But it never becomes dominant among the older age groups.

3.1.1 Scenario modeling

This data set is generated by a simulation that models a hypothetical scenario, where a new variant 60% more transmissible than the circulating variant emerged in Virginia in May 2021. The prior immunity in the population at the beginning of the outbreak period is modeled based on the confirmed cases in Virginia and actual age-stratified vaccination in Virginia prior to May 2021. As mentioned earlier, the scenarios described below are motivated by the scenario model hub (SMH) efforts [1, 3, 4, 2]. We have been participating in the hub activities since its inception in 2021 [12].

Disease model. We use an age-stratified COVID-19 disease model parameterized based on [10], and extended to model two vaccines and two variants. The disease model includes the following states for each age group: S (susceptible), E (exposed), Ipresymp (presymptomatically infectious), Isymp (symptomatically infectious), Iasymp (asymptomatically infectious), rMedAttend (medically attended to be recovered), hMedAttend (medically attended to be hospitalized), dMedAttend (medically attended to be dead), Hosp (hospitalized to be recovered), dHosp (hospitalized to be dead), Vent (ventilated to be recovered), dVent (ventilated to be dead), Death (dead), R (recovered). We extend S state for two vaccines: vaxinitialS (susceptible with initial vaccine) vaxboosterS (susceptible with booster vaccine). An individual vaccinated by the initial vaccine is 90% less susceptible than naively susceptible (90% efficacy against infection). An individual vaccinated by the booster vaccine is 80% less susceptible than naively susceptible (80% efficacy against infection). We also extend the three infectious states for two variants: the existing variant and the new variant. An individual infectious with the new variant is 60% more infectious than with the existing variant. The state transition parameters are presented in Table 1. Note that the extended states have the same transition parameter as their base states thus are omitted from the table.

Initializations. The outbreak is modeled from May 4, 2021 (day 0). It is initialized based on the surveillance data on confirmed cases at county level in Virginia [26] and age specific vaccine administration data of Virginia [11] up to May 4, 2021. From these data sets we derive the distribution of the Virginia population among different compartments: susceptible, vaccinated with initial vaccine, actively infected, and recovered

Progression	Attribute	Age group				
		0-4	5-17	18-49	50-64	65+
E → Iasymp	prob	0.35				
	dt	$N(5,1)$				
Iasymp → R	prob	1				
	dt	$N(5,1)$				
E → Ipresymp	prob	0.65				
	dt	3				
Ipresymp → Isymp	prob	1				
	dt	2				
Isymp → rMedAttend	prob	0.9582	0.9882	0.9582	0.906	0.754
	dt	discrete(1:0.175, 2:0.175, 3:0.1, 4:0.1, 5:0.1, 6:0.1, 7:0.1, 8:0.05, 9:0.05, 10:0.05)				
rMedAttend → R	prob	1				
	dt	$N(5,1)$				
Isymp → dMedAttend	prob	0.0018	0.0018	0.0018	0.009	0.051
	dt-fixed	2				
dMedAttend → dHosp	prob	0.95				
	dt-fixed	2				
dHosp → dVent	prob	0.06	0.06	0.06	0.15	0.225
	dt	2				
dVent → Death	prob	1				
	dt	3				
dHosp → Death	prob	0.94	0.94	0.94	0.85	0.775
	dt	6				
dMedAttend → Death	prob	0.05				
	dt	8				
Isymp → hMedAttend	prob	0.04	0.01	0.04	0.085	0.195
	dt	1				
hMedAttend → Hosp	prob	1				
	dt	$N(5,4.6)$	$N(5,4.6)$	$N(5,4.6)$	$N(5.3,5.2)$	$N(4.2,5.2)$
Hosp → R	prob	0.94	0.94	0.94	0.85	0.775
	dt	$N(3.1,3.7)$	$N(3.1,3.7)$	$N(3.1,3.7)$	$N(7.8,6.3)$	$N(6.5,4.9)$
Hosp → Vent	prob	0.06	0.06	0.06	0.15	0.225
	dt	$N(1,0.2)$				
Vent → R	prob	1				
	dt	$N(2.1,3.7)$	$N(2.1,3.7)$	$N(2.1,3.7)$	$N(6.8,6.3)$	$N(5.5,4.9)$

Table 1: Disease progression parameters. One value per line applies to all age groups. Abbreviations: prob: probability, dt: dwell time, $N(\cdot,\cdot)$: normal distribution, E: exposed, Iasymp: asymptomatically infectious, Ipresymp: presymptomatically infectious, Isymp: symptomatically infectious, rMedAttend: medically attended to be recovered, hMedAttend: medically attended to be hospitalized, dMedAttend: medically attended to be dead, Hosp: hospitalized to be recovered, dHosp: hospitalized to be dead, Vent: ventilated to be recovered, dVent: ventilated to be dead, Death: dead, R: recovered.

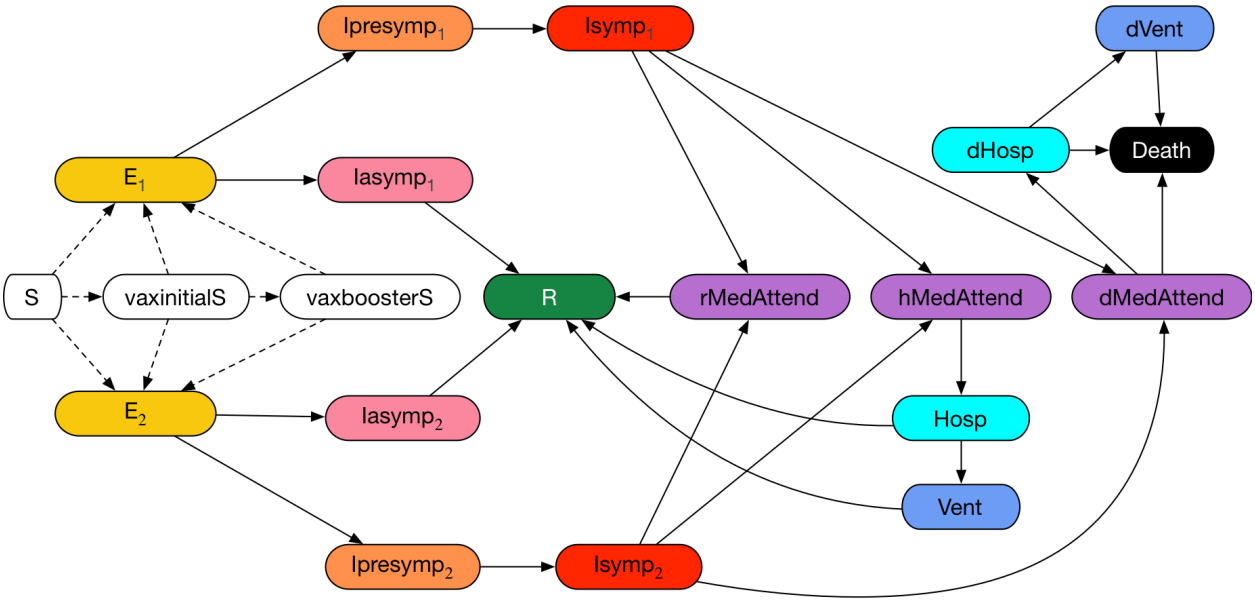


Figure 4: Disease state transition diagram. It has the same structure but different parameterizations (see Table 1) for different age groups. The dashed arrows represent transitions that occur only with external causes (e.g. vaccination or contact with infectious people). The solid arrows represent within-host progressions.

from infection. We use this distribution to assign individuals in our Virginia digital twin to one of these states.

Vaccination with hesitancy. This outbreak models two vaccines: the initial vaccine corresponds to the early two-dose regime of Pfizer and Moderna (and one-dose regime of Johnson&Johnson); the booster vaccine corresponds to the regime introduced in late 2021. The initial vaccine when fully administered provides 90% VE (vaccine efficacy) against infection. The booster provides 80% VE.

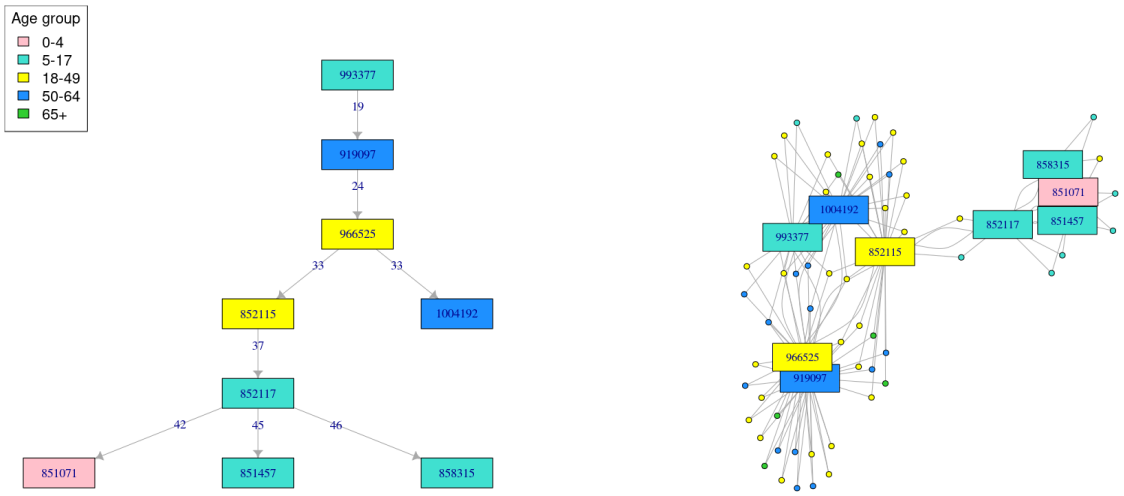
Non-pharmaceutical interventions. After vaccines became available, many of the early stage social distancing measures were relaxed. We model the following non-pharmaceutical interventions.

- School closure: throughout the modeling period, 25% compliance, compliant individuals drop all in-school contacts.
- Mask wearing: throughout the modeling period, 25% compliance, compliant individuals have their transmission risks reduced by 60%.
- Generic social distancing: throughout the modeling period, 15% of all individuals, compliant individuals drop all non-essential activities (shop, religion, other).
- Voluntary home isolation: 14 days from 2 days after becoming pre-symptomatic, 75% of symptomatic cases, compliant individuals drop all non-home activities and contacts.

3.1.2 Highlighted features in the data

In this data set we highlight prior immunity, vaccine prioritization and hesitancy, relaxed NPIs, and competing variants. In both the version delivered to NSF and that to CDC, we include transmissions data, which represent transmission trees as subgraphs of the underlying contact network. It can be used to study e.g. surveillance of an emerging variant. The transmission data of each outbreak consists of a collection of trees, rooted at nodes corresponding to seeded infections (importations in real world). In Figure 5 we show an

example of such trees from an outbreak and the fragment of the contact network that includes the tree nodes and their neighboring nodes. In the delivered data set, we also include transitions data, which records state transition events that have occurred to individuals during the outbreak. The transitions data can be used to study population immunity dynamics at individual level.



(a) A transmission tree from an outbreak. Nodes are labeled by person IDs. Arrows show transmission direction and are labeled by the simulation day the transmissions occur.

(b) Subgraph of contact network corresponding to the tree nodes. It includes the tree nodes (labeled by person IDs) and their neighboring nodes (not labeled for clarity).

Figure 5: Illustration of transmissions data. On the left is one of the transmission trees from an outbreak, which is a random realization of the stochastic diffusion process starting from person with ID 993377 and along the edges of the contact network, of which a small fragment related to the tree nodes is shown on the right.

3.1.3 Data schema

The synthetic population and contact network data sets of Virginia, as well as the synthetic outbreak data sets are delivered in a series of comma-separated-value (CSV) files as described in Table 2. Each file has data for a particular component (e.g., person, household, etc.) of the overall population schema. The fields contained in each file, along with their descriptions, are provided in tables 3 – 10 and figure 6 shows the relationships between different files.

4 Near term applications to Genomics

The interplay between pathogen evolution and societal dynamics presents a significant challenge for genomic epidemiology. Understanding how factors like demographics, mobility patterns, and public health interventions influence the spread and evolution of diseases is crucial for effective surveillance and control. The ability to generate realistic, yet completely synthetic, outbreak data can be further enhanced by introducing evolutionary processes and the generation of synthetic pathogen genomes. Our system, which simulates

Data Component	Description
Person	Each row represents one synthetic individual in the population, including their age, gender, and the household to which they belong.
Household	Each row represents one synthetic household in the population, including its residence location, administrative regions, and the number of household members.
Residence Locations	Each row represents a residence location (e.g., where a household may reside.)
Activity Locations	Each row represents a non-residence location where people may go over the course of the day (e.g., work, school, shopping, etc.)
Activity Location Assignment	Each row maps an individual to an activity and the location where that activity took place. An individual will likely have multiple activity locations over the course of a day.
Population Network	Indicates when and where two people came in contact, and for how long.
Transmissions	Infecting and infected IDs, day, and virus variant of each transmission.
Transitions	The person ID whose state changes, day, and the state after the change of each state transition.

Table 2: Various components of the synthetic population dataset. The naming convention is: {region}_{component}_ver_{major}_{minor}.csv, where {major} and {minor} indicate the version of the population and component indicates person, household, etc.

Field	Description
pid	Person ID: A unique integer identifying a person
hid	Household ID: An integer identifying a household as defined in the Household file
person_number	The sequence identifier related to the indicated person's position within the household. A household with 3 people would have person_numbers 1, 2, and 3.
age	Age of person
sex	Gender of person

Table 3: Person file

Field	Description
hid	Household ID: A unique integer identifying a household
rlid	Residence location ID
admin1	For UK, this is the ADCW ID for the admin1 region; for Virginia, USA, this is the state FIPS code (51)
admin2	For UK, admin2 is the same as admin1 because ADCW does not provide admin2 level for the UK. For Virginia, USA, this is the county FIPS code
hh_size	Household Size: Number of persons in a household

Table 4: Household file

Field	Description
rlid	Residence Location ID: A unique integer identifying the residence location.
longitude	Longitude of the location
latitude	Latitude of the location
admin1	See household file description
admin2	See household file description

Table 5: Residence Locations file

Field	Description
alid	Activity Location ID: Unique integer identifying the location where non-HOME activities can take place
longitude	Longitude of the location
latitude	Latitude of the location
admin1	See household file description
admin2	See household file description
work	Does the location support work activities? (Value is 0 or 1)
shopping	Does the location support shopping activities? (Value is 0 or 1)
school	Does the location support school activities? (Value is 0 or 1)
other	Does the location support other activities? (Value is 0 or 1)
college	Does the location support college activities? (Value is 0 or 1)
religion	Does the location support religion activities? (Value is 0 or 1)

Table 6: Activity Locations file

Field	Description
hid	Household ID of the person
pid	Person ID of the person
activity_number	Activity Number: Number of the activity in the activity sequence to which it belongs
activity_type	Activity Type: Enumerations used for encoding activity types. 1: Home, 2: Work, 3: Shopping, 4: Other, 5: School, 6: College, 7: Religion
start_time	Start time of the activity in seconds since midnight Sunday/Monday
duration	Duration of the activity in seconds
lid	Location ID of the location where the activity takes place (rlid or alid)

Table 7: Activity Location Assignment file

Field	Description
pid1	Person ID 1 of this edge
pid2	Person ID 2 of this edge
lid	Location ID of the location where the contact takes place (rlid or alid)
start_time	Start time of the contact between Person ID 1 and Person ID 2 measured in seconds since midnight of Sunday/Monday
duration	Duration of the contact measured in seconds
activity1	Activity type of Person ID 1 at time of contact, see activity_type in activity location assignment file description
activity2	Activity type of Person ID 2 at time of contact, see activity_type in activity location assignment file description

Table 8: Population (Contact) Network file

Field	Description
day	Simulation day of the transmission
pid	ID of the infected person
variant	Virus variant being transmitted (1: existing variant; 2: new variant)
contact_pid	ID of the infecting person

Table 9: Transmissions

Field	Description
day	Simulation day of the transition
pid	ID of the transitioning person
disease_state	Disease state to which the person transitions

Table 10: Transitions

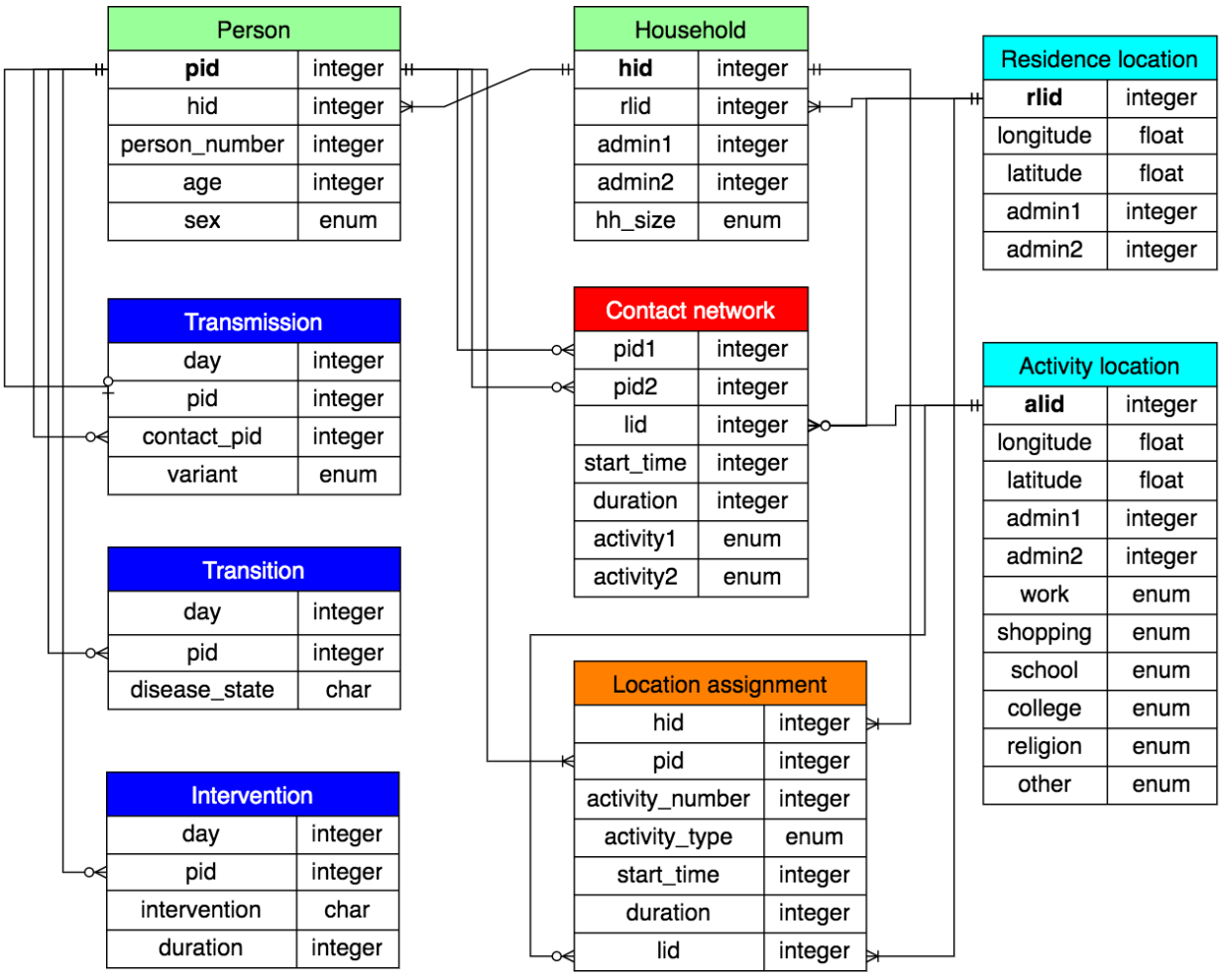


Figure 6: Data structure diagram.

outbreaks on a digital twin of the population, provides a powerful tool for benchmarking genomic epidemiology applications. This approach allows us to explore the combined effects of evolutionary and spreading processes within a realistic societal framework, offering insight into how genomic data can be leveraged for public health analysis.

4.1 Specific applications in Genomic Epidemiology

- **Understanding the Informational Capacity of Pathogen Genomics:** By embedding genomic data within our simulations, we can study the "informational capacity" of pathogen genetic material tuned specifically to the rate and type of change in a specific pathogen, to explore how effectively genomic data can distinguish:
  - **Infection cascades:** Identifying transmission chains and superspreader events.
  - **Demographics:** Understanding how disease spread varies across age groups, socioeconomic strata, and other demographic factors.
  - **Behavioral Groups:** Discerning transmission patterns associated with specific behaviors and risk factors.
- **Epidemiological enhancement:** Digital similars in this context have great potential for capturing a high level of detail and providing situational awareness when combined with real world surveillance
  - **Communication and Situational Awareness:** Creating representations of outbreaks constrained by real world genomic surveillance for analysis and communication purposes. And further, leveraging the digital twin as container for privacy preservation and modulating representation of PHI in an epi context.
  - **Education:** Creating realistic synthetic data sets with matched, detailed demographics and genomic sequences provide the necessary input for demonstrating the full feature set of a range of phylodynamic and bioinformatic tools without an inherent PHI concern.
- **Benchmarking Genomic Epidemiology Applications:** The system provides a controlled environment to develop, test, and benchmark a range of applications, including:
  - **Cascade Reconstruction:** Evaluating the performance of algorithms that reconstruct transmission chains from genomic data, especially under realistic conditions of incomplete sampling and backfill.
  - **Surveillance Sensitivity Analysis:** Generating tailored data sets that mimic real-world surveillance data, allowing researchers to test new analytical methods against features both explicit and emergent.
- **Exploring Specific Research Questions:** The flexibility of our system allows for in-depth investigations of key questions in genomic epidemiology:
  - **Evolutionary Modeling:** Simulating pathogen evolution at a transactional (person-to-person) level to understand how micro-level mutational forces translate into macro-level genomic patterns observed in outbreaks.
  - **Realistic Importation Scenarios:** Modeling diverse importation scenarios to understand how imported cases influence outbreak dynamics and genomic diversity.
  - **Impact of Sampling Strategies:** Examining how different sampling frames and biases (e.g., testing availability, symptom-based testing) affect the observed genomic data and downstream analyses.
  - **Ablation Studies:** Systematically removing or altering specific data components (e.g., mobility data, intervention data) to quantify their impact on genomic inferences and understand the value of different data sources for genomic epidemiology.

By providing a realistic, yet completely synthetic, testing ground, our system enables a deeper understanding of how pathogen genomics interacts with societal factors. This will be crucial for developing robust genomic surveillance systems, designing effective interventions, and ultimately strengthening our ability to combat infectious diseases.

References

[1] The COVID-19 Scenario Modeling Hub. https://covid19scenariomodelinghub.org/.

[2] The European COVID-19 Scenario Hub. https://covid19scenariohub.eu/.

[3] Flu Scenario Modeling Hub. https://fluscenariomodelinghub.org/.

[4] RSV Scenario Modeling Hub. https://github.com/midas-network/rsv-scenario-modeling-hub.

[5] C. Barrett, R. Beckman, M. Khan, V. S. Anil Kumar, M. Marathe, P. Stretz, T. Dutta, and B. Lewis. Generation and analysis of large synthetic social contact networks. In *Winter Simulation Conference*, pages 1003–1014. Winter Simulation Conference, 2009.

[6] C. L. Barrett, R. J. Beckman, M. Khan, V. A. Kumar, M. V. Marathe, P. E. Stretz, T. Dutta, and B. Lewis. Generation and analysis of large synthetic social contact networks. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pages 1003–1014. IEEE, 2009.

[7] R. J. Beckman, K. A. Baggerly, and M. D. McKay. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429, 1996.

[8] P. Bhattacharya, D. Machi, J. Chen, S. Hoops, B. Lewis, H. Mortveit, et al. AI-driven agent-based models to study the role of vaccine acceptance in controlling COVID-19 spread in the US. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1566–1574. IEEE, 2021.

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

[10] CDC. COVID-19 pandemic planning scenarios. https://stacks.cdc.gov/view/cdc/88617, May 2020.

[11] Centers for Disease Control and Prevention. COVID Data Tracker. https://covid.cdc.gov/covid-data-tracker/. Last accessed: March 24, 2023.

[12] J. Chen, P. Bhattacharya, S. Hoops, D. Machi, A. Adiga, H. Mortveit, S. Venkatramanan, B. Lewis, and M. Marathe. Role of heterogeneity: National scale data-driven agent-based modeling for the US COVID-19 Scenario Modeling Hub. *Epidemics*, 48:100779, 2024.

[13] W. E. Deming and F. F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal tables are known. *Annals Math. Stats*, 11(4):427–444, 1940.

[14] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[15] S. Eubank, C. L. Barrett, R. Beckman, K. Bisset, L. Durbeck, C. Kuhlman, B. Lewis, A. Marathe, M. Marathe, and P. Stretz. Detail in network models of epidemiology: Are we there yet? *Journal of Biological Dynamics*, 4:446–455, 2010. PubMed PMID: 20953340; PMCID: PMC2953274.

[16] 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.

[17] S. Eubank, V. S. Anil Kumar, M. V. Marathe, A. Srinivasan, and N. Wang. Structure of Social Contact Networks and Their Impact on Epidemics. In *Discrete Methods in Epidemiology*, volume 70, pages 179–200. American Math. Soc., Providence, RI, 2006.

[18] Federal Committee on Statistical Methodology. STATISTICAL POLICY WORKING PAPER 22 (second version, 2005). Technical Report 22, Office of Management and Budget, Office of Information and Regulatory Affairs, 2005.

[19] G. Harrison, J. Chen, H. Mortveit, S. Hoops, P. Porebski, D. Xie, M. Wilson, P. Bhattacharya, A. Vullikanti, L. Xiong, and M. Marathe. Synthetic data to support us-uk prize challenge for developing privacy enhancing methods: Predicting individual infection risk during a pandemic, 2023.

[20] HERE, 2020. http://www.here.com, Accessed April 2020.

[21] 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.

[22] Microsoft. U.S. building footprints. https://github.com/Microsoft/USBuildingFootprints, 2020.

[23] H. S. Mortveit, A. Adiga, C. L. Barrett, J. Chen, Y. Chungbaek, S. Eubank, C. J. Kuhlman, B. Lewis, S. Swarup, and S. Venkatramanan. Synthetic populations and interaction networks for the U.S. Technical Report 2019-025, NSSAC, University of Virginia, 2020.

[24] T. National Center for Education Statistics (NCES). Last accessed: February 2020.

[25] S. Swarup and M. Marathe. Generating synthetic populations for social modeling. AAMAS, 2017.

[26] The New York Times. Coronavirus (covid-19) data in the United States. https://github.com/nytimes/covid-19-data. Last accessed: March 24, 2023.

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

[28] United States Censuc Bureau. 2011-2015 5-year ACS commuting flows. Last accessed: April 2020.

[29] United States Census Bureau. American Community Survey 2013-2017 5-year estimates. Last accessed: February 2020.

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

[31] US Census Bureau. Disclosure avoidance for the 2020 census: An introduction. Technical report, US Government Publishing Office, 2021.

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