

1 Synthetic Population Generation with Public Health Characteristics for Spatial Agent-Based Models

2 Emma Von Hoene^{1*}, Amira Roess², Hamdi Kavak³, Taylor Anderson¹

3 1. Department of Geography and Geoinformation Science, College of Science, George Mason
4 University, Fairfax, VA, USA

5 2. Department of Global and Community Health and Epidemiology, College of Public Health,
6 George Mason University, Fairfax, VA, USA

7 3. Department of Computational and Data Sciences, College of Science, George Mason
8 University, Fairfax, VA, USA

9 *corresponding author

10 E-mail: evonhoen@gmu.edu

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28 **Synthetic Population Generation with Public Health Characteristics for Spatial Agent-Based**

29 **Models**

30 **Abstract**

31 Agent-based models (ABMs) simulate the behaviors, interactions, and disease transmission
32 between individual “agents” within their environment, enabling the investigation of the underlying
33 processes driving disease dynamics and how these processes may be influenced by policy
34 interventions. Despite the critical role that characteristics such as health attitudes and vaccination
35 status play in disease outcomes, the initialization of agent populations with these variables is often
36 oversimplified, overlooking statistical relationships between attitudes and other characteristics or
37 lacking spatial heterogeneity. Leveraging population synthesis methods to create populations with
38 realistic health attitudes and protective behaviors for spatial ABMs has yet to be fully explored.
39 Therefore, this study introduces a novel application for generating synthetic populations with
40 protective behaviors and associated attitudes using public health surveys instead of traditional
41 individual-level survey datasets from the census. We test our approach using two different public
42 health surveys (one national and the other representative of the study area, Virginia, U.S.) to create
43 two synthetic populations representing individuals aged 18 and over in Virginia, U.S., and their
44 COVID-19 vaccine attitudes and uptake as of December 2021. Results show that integrating public
45 health surveys into synthetic population generation processes preserves the statistical relationships
46 between vaccine uptake and attitudes in different demographic groups while capturing spatial
47 heterogeneity at fine scales. This approach can support disease simulations that aim to explore how
48 real populations might respond to interventions and how these responses may lead to demographic
49 or geographic health disparities. Our study also demonstrates the potential for initializing agents with
50 variables relevant to public health domains that extend beyond infectious diseases, ultimately
51 advancing data-driven ABMs for geographically targeted decision-making.

52 **Author Summary**

53 In this study, we introduce a new method for generating synthetic populations of individuals or
54 “agents” with characteristics that include health protective behaviors and attitudes, which are
55 crucial for modeling disease spread. Traditional methods for parameterizing agents often overlook
56 the complex relationships between demographic factors and health behaviors like vaccination.
57 Additionally, detailed spatial data capturing these behaviors are limited, meaning agent behaviors
58 are more uniform across geographic space. By fitting public health surveys with spatially aggregated
59 census data, we created more realistic agent populations for disease spread simulations. We
60 focused on Virginia, U.S. and generated a population with COVID-19 vaccine uptake and attitudes as
61 of December 2021. Our results show that this approach captures the statistical relationships
62 between demographic variables and vaccine uptake, along with the spatial variation in these
63 behaviors. We also show that using national survey data is comparable to using local survey data
64 representative of Virginia collected in 2021. The approach is flexible so that it can be applied to
65 various public health studies beyond just infectious diseases. Our work highlights the potential of
66 public health surveys for enhancing synthetic population generation, offering a valuable approach
67 for initializing models with more realistic populations to explore public health challenges.

68 **Keywords:** synthetic population generation; agent-based models; public health applications;
69 infectious disease simulations; vaccine uptake; vaccine attitude; COVID-19

70 **1. Introduction**

71 Agent-based models (ABMs) are commonly used to simulate the spread of infectious diseases
72 caused by viruses, including COVID-19 virus (1–4), influenza (5,6), and the chickenpox virus (7,8).
73 Unlike traditional epidemiological models, such as the Susceptible-Infectious-Recovered (SIR)
74 model and its variants, ABMs use a bottom-up approach that simulates the behaviors, interactions,

75 and subsequent transmission of disease between individual “agents” within their environment
76 (9,10). This approach allows for the investigation of the underlying processes driving disease
77 dynamics and how these processes may be influenced by policy interventions (11).

78 Given the important role of demographic characteristics such as age and income (12), household
79 structures (13–15), activity patterns and co-location (16) in disease dynamics, most ABMs of disease
80 spread attempt to incorporate these attributes when initializing agents. For example, children often
81 participate in activities like attending school or recreational events, where they interact with many
82 other individuals and are more likely to contract pathogens that can then be transmitted to parents
83 or grandparents living in the same household (17). While some studies use random functions or fixed
84 values to assign agent attributes, population synthesis approaches can utilize spatially aggregated
85 census data and individual-level survey data (18) from sources like household travel surveys (19) or
86 census microdata (20) to create a complete agent population with relevant attributes, including
87 household structures, thus accurately capturing these transmission pathways within the model.

88 Disease dynamics are also shaped by the uptake of protective behaviors by the population, such as
89 wearing masks, getting vaccinated, and staying home when sick, which can reduce the likelihood of
90 negative health outcomes (21). In addition to social norms and physical or financial barriers, an
91 individual’s attitudes, beliefs, and perceptions significantly affect their decision to engage in
92 protective behaviors. Although traditionally overlooked in ABMs of disease spread (22–24), the
93 COVID-19 pandemic spurred on a widespread effort to better represent health behaviors and their
94 dynamics into epidemiological models (25–28). This paper argues that a synthetic population
95 generation approach capable of initializing agent populations with a realistic set of attitudes and
96 protective behaviors can support such ABMs that aim to simulate behavior dynamics influenced by
97 these attributes.

98 A typical approach in current ABMs is that protective behaviors or related attitudes are assigned to
99 agents with some probability, either based on a hypothetical scenario or based on aggregate data
100 measuring the real characteristics of the population. For example, Rafferty et al. (7) use an ABM to
101 simulate the impact of dose timing, coverage, and waning of immunity on chickenpox disease
102 outcomes in Alberta, Canada. They initialize the population with vaccination attitudes based on
103 aggregate data (65% acceptance, 30% hesitant, 5% reject). However, this approach ignores the
104 statistical relationships between vaccine attitudes and other individual demographic, cultural, or
105 political characteristics, that synthetic population generation approaches aim to preserve.

106 In another example, Pandey et al. (29) use an ABM to examine the effect of bivalent boosters on
107 COVID-19 outcomes, assuming a coverage of 59%, 51%, 38%, 54%, and 75% for age groups 5-11, 12-
108 17, 18-49, 50-64 and 65+, respectively, informed by historical influenza data. While their model more
109 accurately captures the relationship between booster coverage and age, whereby age 65+ are more
110 likely to accept a booster, the study assumes spatially uniform uptake across New York City. This
111 assumption of uniformity is common, especially since health data are often not available at
112 granularities finer than county or state, meaning that spatial heterogeneities can only be captured at
113 these levels. While numerous ABMs have been developed to simulate the adoption of protective
114 behaviors or the spread of beliefs, attitudes, perceptions towards vaccines over space and time, the
115 use of synthetic population approaches to initialize an agent population with these characteristics
116 has yet to be explored.

117 Therefore, the purpose of this study is to investigate how synthetic population generation approaches
118 can be expanded to create agent populations with attitudes and initial adoption of protective
119 behaviors, along with their spatial distributions. Specifically, we aim to replace datasets commonly
120 used in synthetic population generation that provide individual-level data from samples with coarser
121 geographic resolution, such as the Census Bureau's Public Use Microdata Sample (PUMS) (20), with

122 public health surveys. Using COVID-19 as a case study, we explore the potential for this approach by
123 generating a synthetic population representing Virginia, U.S. and their vaccine attitudes and uptake
124 as of December 2021. We obtain real vaccine uptake for Virginia at the CT level for the same point in
125 time to validate our results, comparing the populations generated by two different public health
126 surveys: one national, and the other representative of Virginia.

127 **2. Background**

128 With the growing use of ABMs across disciplines such as economics, geography and biology (30), a
129 wealth of synthetic population generation methods have been developed to create agent
130 populations. These populations serve as simplified microscopic representations of the targeted
131 population, reflecting individuals and their socio-demographic characteristics relevant to the study
132 (31). The emergence of synthetic population generation approaches is largely due to several factors,
133 including privacy restrictions that prevent access to detailed individual-level data at fine spatial
134 scales, the ability of ABMs to simulate social dynamics and behaviors which are connected with
135 individual attributes, and advancements that have made ABMs more data-informed and effective as
136 predictive tools for decision support (32,33).

137 Synthetic population generation methods vary in complexity and are broadly categorized into two
138 main approaches: Combinatorial Optimization (CO) and Synthetic Reconstruction (SR). CO focuses
139 on replicating real entities by reweighting an existing dataset to match individual profiles. In contrast,
140 SR, which is more commonly used and well-established, generates populations through random
141 sampling from known distributions of demographic characteristics or estimated joint distributions
142 using deterministic re-weighting algorithms like Iterative Proportional Fitting (IPF) (31,34). Given the
143 extensive literature on population synthesis, we provide only a brief background to support

144 understanding of our proposed method. For a comprehensive review of population generation
145 approaches for ABMs, see Chapuis et al. (31).

146 IPF is the most widely used approach for generating synthetic populations due to its long-standing
147 presence and reliability in literature, computational efficiency, and its methodological simplicity (35).
148 The algorithm adjusts each cell in an n-dimensional matrix, which represents the distribution of
149 attributes, based on known marginal controls. It starts with sample data to initialize the matrix and
150 then iteratively updates the cells to match the specified contingency dimensions (36). Originally
151 introduced by Deming and Stephan (37) to adjust contingency tables to fit with known marginal
152 distributions, IPF has been extensively refined by researchers to improve its application for
153 population synthesis. For instance, Beckman et al. (38) first established the methods for using IPF
154 with PUMS data, where joint distributions of household attributes were derived by integrating sample
155 frequency tables from PUMS data with marginal distributions from Census Summary Files, and then
156 randomly selecting households based on these estimates to create a synthetic population.

157 Synthetic population generation approaches, such as those using IPF to initialize an agent population
158 within a spatial ABM, typically combine spatially aggregate and disaggregated individual-level data
159 to statistically match both the joint distributions found in the individual-level data with the marginal
160 totals in the spatially aggregate data (18). Spatially aggregate data captures marginal totals of
161 populations across a set of categories such as gender, age, and race within different geographic
162 zones (e.g. census tracts, dissemination areas) in a study area. This data allows for analysis of
163 populations and their spatial distributions with relatively fine granularity while preserving privacy by
164 presenting only marginal totals (e.g., total population aged 65+, or total population that is white)
165 rather than joint distributions across multiple attributes (e.g., total population aged 65+ and white).

166 Disaggregated individual-level survey data contains samples of anonymized records of real
167 individuals and their demographic characteristics. Although this data captures the joint distributions

168 among individual attributes, it represents only a small sample from a large geographic area (e.g., a
169 state or the entire country), which protects individual identities and prevents inference of the spatial
170 distribution of the sample population. Examples of such datasets include PUMS in the U.S., and
171 similar datasets available in other countries, such as Public Use Microdata Files (PUMFs) in Canada.

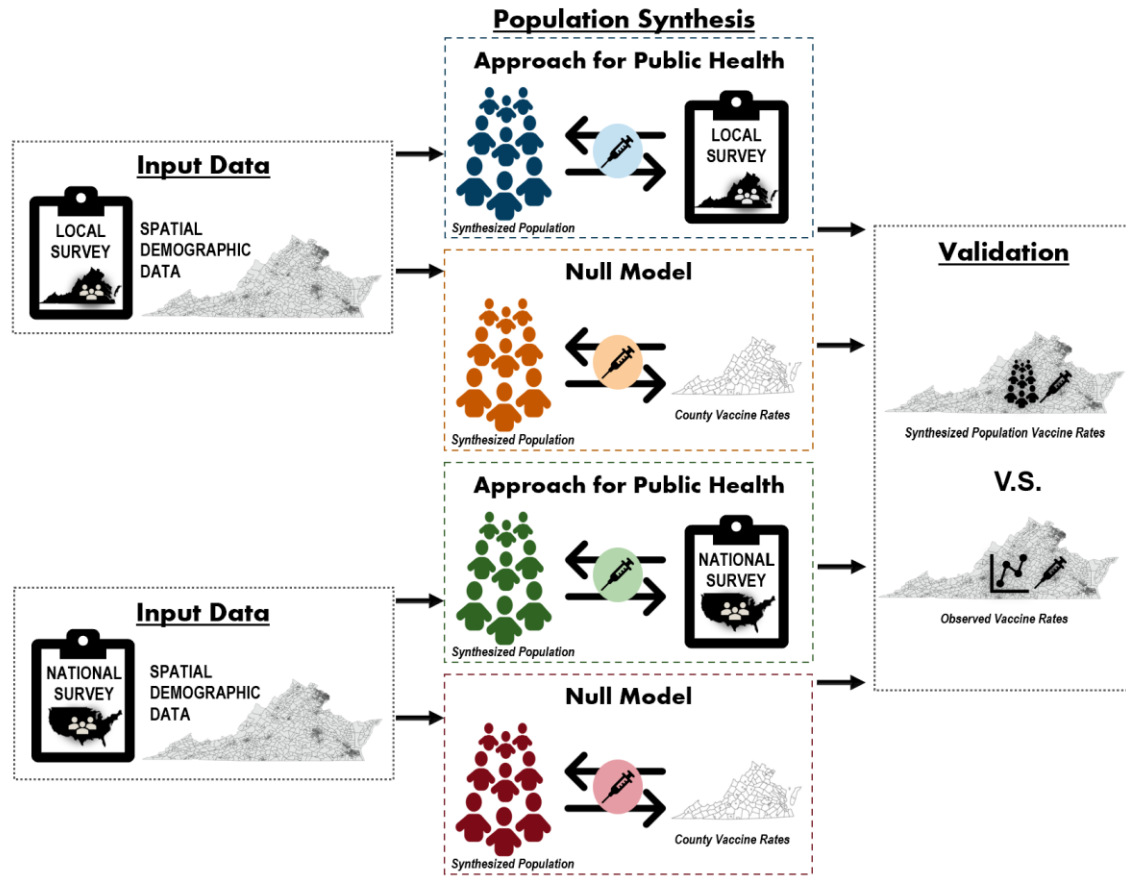
172 Synthetic population generation, particularly SR approaches, involves both fitting and allocation.
173 During the fitting stage, IPF is used to align individual-level sociodemographic data with spatially
174 aggregated constraints, generating fractional weights for entities such as households or individuals
175 in each geographic zone. Because IPF outputs fractional weights, allocation is required to produce a
176 discrete set of agent counts that replicates individuals (32). The fractional weights are converted into
177 integer weights through a process known as ‘integerisation’, which can be performed using various
178 approaches such as simple rounding, thresholding, proportional probabilities, or truncate, replicate,
179 sample (TRS) (see Lovelace and Ballas (39) for a review on these methods). ‘Integerisation’ is
180 followed by expansion, where each individual is represented as a record with a geographic zone, and
181 the matching attributes for that individual from the original survey dataset is carried over (39).

182 IPF has been used to create agent populations in various spatial ABMs, such as those for disaster
183 management and recovery (40), though it is most commonly used in urban and transportation
184 modeling. However, in the context of spatial ABMs for infectious disease spread, there are few
185 dedicated population synthesis methods or studies utilizing well-established techniques such as IPF
186 (41). This is likely because ABMs take significant time to develop and are often designed for specific
187 objectives, such as understanding the impact of policy guidelines and health behaviors on infectious
188 disease dynamics (42–44), proposing general or behavioral frameworks for epidemiological models
189 (45,46), or forecasting disease transmission (48,49). This gap is particularly significant as these
190 models are valuable for informing policy, yet generating populations with detailed individual
191 characteristics and health behaviors often remains overlooked, despite their critical role in

192 influencing disease transmission. To our knowledge, no specific synthetic population generation
193 method for spatial ABMs of disease spread has yet been developed to capture both individual
194 attitudes and initial adoption of protective behaviors, along with the spatial heterogeneities in these
195 characteristics. Therefore, there is a need for a flexible synthetic population generation approach
196 that accurately initializes agent populations with attitudes, beliefs, perceptions and initial adoption
197 of protective behaviors, as well their spatial distributions. By proposing a targeted population
198 synthesis method that derives these individual attributes from public health surveys, this approach
199 can be adapted for various public health applications, including infectious diseases, smoking, and
200 other health challenges, across different scales and locations.

201 **3. Materials and Methods**

202 Our approach is presented in Figure 1. First, individual-level survey data and spatially aggregate data
203 are used as input data for the population synthesis of agents with demographic characteristics. Our
204 approach extends traditional synthetic population generation approaches by allowing for
205 vaccination status and attitudes to be carried over at the replication stage. We compare our
206 approach with a null model, which uniformly assigns vaccine uptake likelihood based on county level
207 vaccine uptake data. Our validation involves comparing the vaccination rates in the synthetic
208 population to those in the real population for each census tract. This comparison is conducted for
209 populations generated using two different public health surveys (a local survey and a national survey)
210 and their respective null models. The data and the methods are described in detail in the following
211 sections. The code written in the R scripting language and the data for the synthetic population
212 generation approach and the validation is available at the GitHub repository
213 <https://github.com/evonhoene/Population-Generation-for-Public-Health-ABMs>.



214

215

Figure 1. Overview of the approach used in the study.

216

3.1. Input Data

217 Given that IPF is a well-established, efficient, and straightforward method for synthetic population

218 generation, we use it to ensure the flexibility of our proposed approach for various study applications.

219 This method requires both spatially aggregated demographic data and individual-level survey data.

220 For the spatially aggregated data, we use census tract (CT) data from the Census Bureau's American

221 Community Survey (ACS) (50) that captures marginal totals for sociodemographic variables. We

222 focus on gender, race, age, education, and income variables for individuals aged 18 and over, as

223 these factors significantly influence COVID-19 vaccine uptake (51). While the ACS provides marginal

224 totals for individuals across different categories of gender, race, age, and education, income data is

225 reported as the percentage of households within each CT that fall into specific income brackets. To

226 estimate individual income levels, we calculate the proportion of the total population aged 18+ that
227 would fall into each income bracket, assuming a household size of 1. We use 2021 data specifically
228 for Virginia CTs and exclude records with missing or zero values for any variable, resulting in a final
229 dataset of N = 2162. Descriptive statistics for the variables collected from the ACS dataset are
230 presented in Table 1.

231 **Table 1. Descriptive statistics summarizing the demographic distribution within Virginia Census Tracts**

Variable: Descriptor	Mean %	Minimum %	Median %	Maximum %
Gender: Male	48.70285	0	48.73321	100
Gender: Female	51.29715	0	51.26679	100
Race/Ethnicity: White	63.14219	0	66.38809	100
Race/Ethnicity: Black	19.90094	0	12.77431	100
Race/Ethnicity: Hispanic	7.914201	0	5.032508	80.2409
Race/Ethnicity: Other	9.065319	0	5.853839	62.07447
Age: 18-29	20.41858	0	17.94349	98.78631
Age: 30-49	33.61571	0	33.13478	71.48159
Age: 50-64	25.43375	0	25.9862	100
Age: 65 and over	20.53196	0	19.91834	100
Education: Bachelor's degree or higher	17.29956	0	14.93085	53.30806
Education: No Bachelor's degree	82.70044	46.69194	85.06915	100
Income: Less than \$25,000	14.95912	0	11.69449	100
Income: \$25,000 - \$49,999	17.3773	0	17.09721	63.68978
Income: \$50,000 - \$99,999	28.74916	0	28.9908	100
Income: Greater than \$100,000	39.18442	0	35.24403	100

232
233 Our approach replaces the traditional individual-level samples captured by censuses used in
234 synthetic population generation (e.g. the PUMS in the US) with public health surveys. We compare
235 the results of our approach using two surveys, one that is one that is representative of Virginia and
236 one that is national and publicly available, as follows:

237 1) *Local Survey*: This survey, collected by researchers is representative of the Commonwealth
238 of Virginia and includes data on demographics as well as beliefs, attitudes, and perceptions
239 related to COVID-19 and protective behaviors. The sample was recruited by Climate Nexus
240 Polling (August 15-31, 2021), using several market research panels. Participants were

241 recruited using stratified sampling methods. Compensation for participants depended on
242 the specific market research panel and respondents' preferences (e.g., cash, gift cards,
243 reward points). Sampling weights accounted for small deviations from the pre-selected
244 census parameters. The dataset includes N = 3,528 respondents. The descriptive statistics
245 for the data are provided in Table 2. De-identified data are available upon request. This
246 project to collect the local survey data was considered exempt by the George Mason
247 University IRB (IRB 1684418-3).

248 2) *Household Pulse Survey (HPS)*: This publicly available national survey, obtained from the US
249 Census Bureau (52), measures the impact of emergent social and economic issues on
250 households across the country, including COVID-19 vaccinations. The HPS also collects data
251 on core demographic characteristics from respondents aged 18 and older. We use data from
252 HPS Week 41, covering December 29, 2021, to January 10, 2022. Records missing data for
253 one or more variables were removed (e.g., vaccine decision, household income), resulting in
254 a total of N = 63,180 respondents. Given the large size of the HPS dataset, we use stratified
255 sampling to reduce the sample to 3,500 to match the size of the *local* survey. As described in
256 Table 2, the HPS data shows a bias, with 91.19% of respondents reporting being vaccinated.
257 At the same time, publicly available county-level vaccination data from the CDC (53)
258 indicates that only 50.2% of Virginians were vaccinated by December 30, 2021. To correct
259 this bias in our stratified sample of 3,500 records, we adjust so that 50% of the sample is
260 vaccinated while preserving the representation of all other variables.

261

262

263

264 **Table 2. Comparative distribution of respondent characteristics from the individual-level surveys**

Variable: Descriptor	% Respondents from HPS	% Respondents from Local Survey
Gender: Male	41.28	44.76
Gender: Female	58.72	55.24
Race/Ethnicity: White	75.06	74.12
Race/Ethnicity: Black	7.01	10.15
Race/Ethnicity: Hispanic	9.16	10.86
Race/Ethnicity: Other	8.78	4.88
Age: 18-29	9.15	16.10
Age: 30-49	36.93	39.12
Age: 50-64	29.15	20.29
Age: 65 and over	24.76	24.49
Education: Bachelor’s degree or higher	41.86	36.96
Education: No Bachelor’s degree	58.14	63.04
Income: Less than \$25,000	11.74	52.15
Income: \$25,000 - \$49,999	19.72	
Income: \$50,000 - \$99,999	31.03	28.66
Income: Greater than \$100,000	37.51	19.19
Vaccination: Yes	91.19	65.33
Vaccination: No	8.81	34.67

265

266 Each survey is used to generate a separate set of synthetic population. In the surveys, while some of

267 the individual-level data is measured on a continuous scale (e.g. age), other data are measured

268 categorically, which results in varying levels of detail between the two synthetic populations,

269 depending on the questions asked. For example, when asking about income, the local survey allows

270 respondents to select <\$50,000, \$50,000-\$99,999, and >\$100,000. On the other hand, the HPS

271 allows respondents to select <\$25,000 and \$25,000-\$49,999, \$50,000-\$99,999, and >\$100,000,

272 allowing for slightly more detailed agent characteristics related to income. In any case, to be used in

273 the IPF process, the data measured by the individual-level survey must be able to be fall under the

274 categories in the spatially aggregated ACS data. This was possible for attributes including gender,

275 race and ethnicity, age, education, and income. Descriptive statistics for the variables captured by

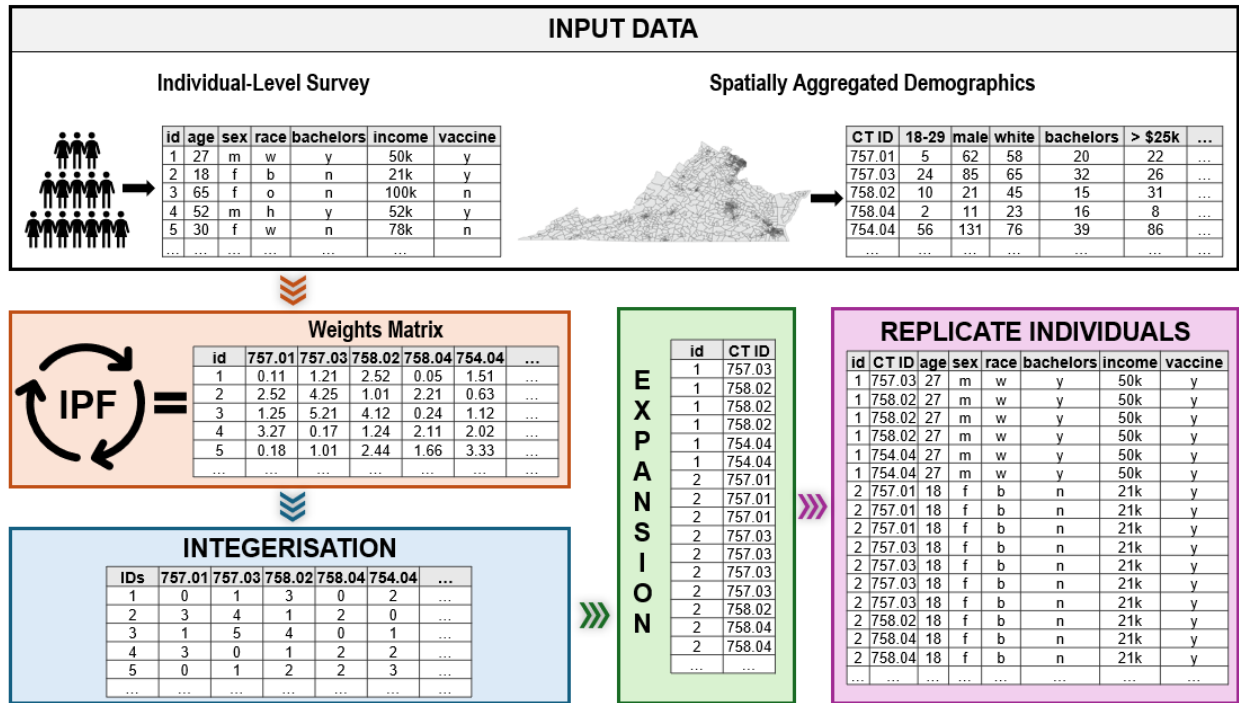
276 the individual-level surveys and their categories from the two survey datasets are outlined in Table 2.

277 Our validation approach uses CT data capturing marginal totals for real vaccine uptake aged 12+ in
278 Virginia as of December 30, 2021. This data is not publicly available and was acquired by request
279 from the Virginia Department of Public Health. The department has since scaled down its operations,
280 and this data is no longer accessible, even upon request. The vaccine uptake data is available for
281 1,601 CTs for which we generate a population. Furthermore, 9 records indicated that vaccine uptake
282 was greater than 100% and were removed. As such our validation only focuses on CTs where data is
283 available, and vaccine uptake is less than or equal to 100% (N=1592).

284 **3.2. Population Synthesis**

285 ***Synthetic population generation approach for public health.*** We use IPF to generate
286 approximately 6 million agents representing the population of Virginia aged 18 and over. This includes
287 generating one population based on the HPS and another using the local survey. Our approach is
288 detailed in Figure 2. IPF computes a weight for every individual in the survey based on how well their
289 characteristics represents the *age, gender, race, income, and education* distributions found in the CT
290 population. These weights are then processed using the TRS ‘integerisation’ method (39), which
291 involves truncating all weights to integers and using these as the counts of each individual type in the
292 zone, followed by sampling to achieve the correct population size based on the probabilities
293 corresponding to the decimal weights. Simply, this approach converts the weights to integers that
294 describe how many times that individual respondent in the survey should be replicated as an agent
295 in the CT. This process is repeated for each CT in the study area. Following this, expansion is
296 conducted to create the final dataset, where each record corresponds to an individual and their CT.
297 By replacing the PUMS with the public health surveys, the demographic variables and every other
298 variable captured by the surveys including a COVID-19 vaccination status as well as attitudes,
299 perceptions, and beliefs related to vaccine are carried over in the sampling and replication stage.

300



301

302

Figure 2. Overview of the synthetic population generation approach for public health.

303

304

305 **Null model.** We compare the results of our public health synthetic population generation approach

306 with a null model that serves as a baseline. Two distinct populations were generated using the null

307 model, corresponding to the HPS and local survey datasets. With the null model, the IPF method fits

308 the individual level demographic data from the surveys with the CT data, creating a population of

309 agents with age, gender, race, income, and education characteristics for each CT in Virginia.

310 However, since vaccine uptake information is only publicly available at county-level, the null model

311 uses this data to impose vaccination uniformly on agents in the same county. For example, as of

312 December 30, 2021, 84.5% of individuals aged 18+ living in Fairfax County were vaccinated (53).

313 Therefore, all agents generated for Fairfax County in the null model are assigned a vaccination

314 likelihood of 84.5%. This is a common approach in ABM to initialize agents with health variables such
315 as vaccine uptake.

316 **Validation.** We compare the spatial and statistical patterns of the simulated vaccine uptake with the
317 observed vaccine uptake percentage at the CT level and with the individual level survey data for the
318 same time period. Although the population is generated for all CTs in the study area, validation is only
319 possible for CTs where real vaccine uptake data is available and where vaccine uptake is less than or
320 equal to 100%.

321 **4. Results**

322 We evaluate the observed and simulated percentages of gender, race, age, education, income, and
323 vaccine status variables for Virginia CTs (N = 1,592) in the populations generated using the HPS and
324 the local survey, using the following quantitative measures: Pearson's correlation coefficient (r),
325 coefficient of determination (r^2), root mean squared error (RMSE), and mean absolute error (MAE).

326 Pearson's correlation coefficient measures the strength and direction of the linear relationship
327 between two variables. The coefficient of determination is the square of this coefficient, providing a
328 quantitative measure of how well one variable explains another. This metric ranges from 0 to 1, where
329 1 represents a perfect fit and values near 0 indicate little to no association. In this context, r^2 assesses
330 how closely the simulated individual characteristics, aggregated by census tract, align with the
331 actual census tract demographic data. Since the IPF approach is designed to fit individual-level data
332 to the marginal totals in census tract data, it is unsurprising that the values of r and r^2 for gender, race,
333 age, education, and income are very high for both surveys. Because vaccine uptake is typically
334 unavailable at the census tract level and cannot be directly incorporated into the IPF, our approach
335 "carries over" individual vaccine status along with their attitudes, beliefs, and perceptions during the
336 sampling and replication stage (see Figure 2).

337 We find that combining IPF with either of the public health surveys allows us to initialize agents with
338 COVID-19 vaccination status in a way that reasonably reflects the real population. The Pearson
339 correlation coefficient and the coefficient of determination for vaccine uptake evaluating the
340 populations generated from both surveys are moderately high (see Table 3). This is visually depicted
341 in Figure 3, where each scatterplot point represents one of the 1,592 Virginia census tracts, with the
342 x-axis showing the observed percentage of vaccine uptake and the y-axis showing the percentage
343 within the synthesized population. In general, within the simulated population using the HPS with
344 IPF, census tracts with higher real vaccination rates also show higher proportions of vaccinated
345 synthetic individuals, with a moderate positive correlation ($r = 0.75$, $r^2 = 0.56$, Figure 3A). A similar
346 pattern emerges in the simulated population from the local survey ($r = 0.72$, $r^2 = 0.51$, Figure 3B).
347 Additionally, when comparing the count of simulated vaccinated individuals in each census tract to
348 the actual count, we find a stronger positive correlation for the HPS dataset ($r = 0.91$, $r^2 = 0.83$, Figure
349 3C) and the local survey dataset ($r = 0.88$, $r^2 = 0.77$, Figure 3D). However, this is largely a reflection of
350 how well the IPF simulates the total population in each census tract, as larger populations naturally
351 lead to more vaccinated individuals.

352 RMSE, measured in the same units as the original data, indicates how closely a simulated population
353 matches the actual census tract (CT) data, with lower values reflecting a better fit and higher values
354 signaling greater discrepancies. As expected, the RMSE values are low for gender, race, age,
355 education, and income. However, the RMSE for the observed and simulated percentage vaccination
356 rates across CTs is 18.28 for the HPS dataset and 13.28 for the local survey dataset. These values
357 suggest that, on average, the simulated percentage of vaccinated individuals differs from the actual
358 percentage by 18.28% and 13.28%, indicating a moderate level of inaccuracy.

359 Similarly, MAE measures the average magnitude of errors between predicted and observed values by
360 averaging the absolute differences, without considering their direction. Unlike RMSE, MAE does not

361 square the errors, making it less sensitive to large deviations and more robust to outliers. MAE values
362 are consistently low for gender, race, age, education, and income variables in both synthetic
363 populations, as IPF effectively fitted these variables to the CT data. For vaccine uptake percentages,
364 MAE values for the synthesized populations are 15.70 for the HPS dataset and 10.65 for the local
365 dataset, which are comparable to the RMSE values. This indicates that the simulated vaccine uptake
366 percentages differ from actual values by 15.70% and 10.65%, respectively, and the similarity
367 between MAE and RMSE values suggests that large deviations do not disproportionately impact the
368 average error. Overall, both RMSE and MAE suggest that the simulated vaccine uptake percentages
369 from our synthetic population generation approach are reasonably close to the observed values
370 across Virginia census tracts. Furthermore, the RMSE and the MAE are smaller for the local survey
371 dataset, possibly since the dataset is representative of Virginia rather than a national dataset. In
372 general, the synthetic census tracts tend to have a smaller proportion of vaccinated individuals than
373 compared to the real population. This may be attributed to the fact that the validation dataset
374 captures vaccination age 12+ and we simulate agents age 18+.

375 In contrast, the null model shows significantly poorer performance in initializing agents realistically
376 with vaccine decisions, as evidenced by a Pearson correlation coefficient of 0.298 and a coefficient
377 of determination of 0.089. These low values indicate a weak relationship between the simulated and
378 observed vaccine uptake percentages. Additionally, the null model's RMSE of 30.357 and MAE of
379 24.008 are considerably higher compared to our proposed approach. These error metrics suggest
380 greater deviations between the simulated and actual vaccine uptake data in CTs, demonstrating that
381 the null model fails to accurately reflect the real distribution of vaccine uptake. This comparison
382 highlights the limitations of the null model in capturing vaccination behaviors when initializing an
383 agent population and emphasizes the improved performance of our synthetic population generation
384 approach using public health surveys.

385

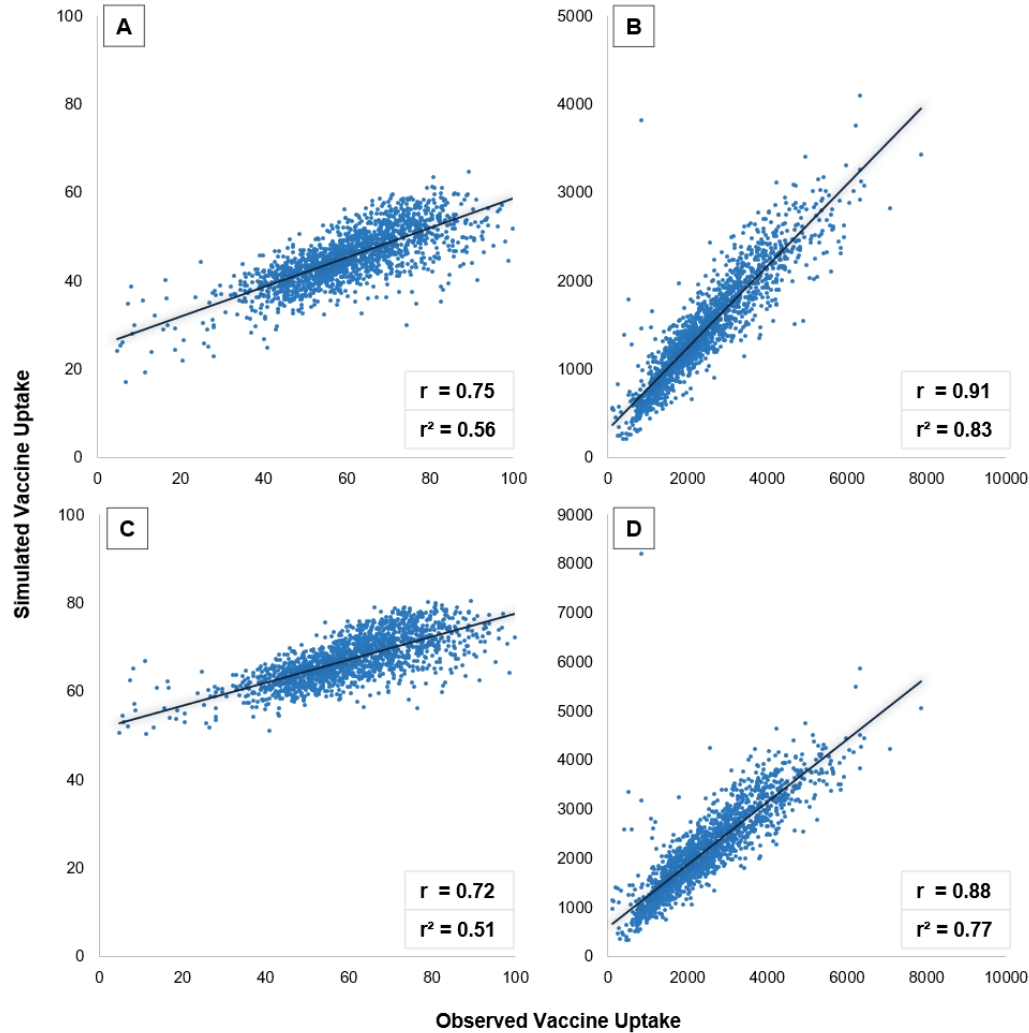
386 **Table 3. Evaluation metrics comparing observed and simulated percentages of demographic and vaccine status**
 387 **variables across Virginia Census Tracts**

Variable: Descriptor	HPS Dataset				Local Survey Dataset			
	Pearson's <i>r</i>	<i>r</i> ²	RMSE	MAE	Pearson's <i>r</i>	<i>r</i> ²	RMSE	MAE
Gender: Male	0.9899	0.9799	0.7784	0.5867	0.9855	0.9712	0.9450	0.7307
Gender: Female	0.9899	0.9799	0.7784	0.5867	0.9855	0.9712	0.9450	0.7307
Race/Ethnicity: White	0.9991	0.9981	1.3229	1.0064	0.9985	0.9969	1.5779	1.2491
Race/Ethnicity: Black	0.9993	0.9987	1.0111	0.6690	0.9990	0.9980	1.1379	0.8145
Race/Ethnicity: Hispanic	0.999	0.9980	0.4696	0.3461	0.9989	0.9978	0.4965	0.3769
Race/Ethnicity: Other	0.9982	0.9965	0.9257	0.5872	0.9985	0.9969	0.5381	0.3612
Age: 18-29	0.9982	0.9963	0.758	0.5076	0.9830	0.9663	2.9901	2.3066
Age: 30-49	0.9948	0.9896	0.9428	0.7392	0.9625	0.9264	3.4143	2.5824
Age: 50-64	0.9945	0.9891	0.7982	0.6127	0.9927	0.9854	0.8424	0.6546
Age: 65 and over	0.9928	0.9857	1.2256	0.9309	0.9963	0.9926	0.7648	0.5728
Education: Bachelor's degree or higher	0.9969	0.9938	1.7922	1.5619	0.9971	0.9941	2.1185	1.8389
Education: No Bachelor's degree	0.9969	0.9938	1.7922	1.5619	0.9971	0.9941	2.1185	1.8389
Income: Less than \$25,000	0.9992	0.9983	0.5555	0.4177	0.9984	0.9967	1.3171	1.0332
Income: \$25,000 - \$49,999	0.9984	0.9967	0.6097	0.4674				
Income: \$50,000 - \$99,999	0.9974	0.9948	0.7428	0.5484				
Income: Greater than \$100,000	0.9994	0.9988	1.0354	0.7246				
Vaccine Uptake: Received	0.7489	0.5608	18.2780	15.6970	0.7151	0.5113	13.2810	10.6457

388

389

390



391

392 **Figure 3. Scatterplots comparing observed and simulated vaccine uptake within Virginia Census Tracts: A)**
393 **percentage from the HPS, B) count from the HPS, C) percentage from the local survey, D) count from the local**
394 **survey.**

395 Our approach effectively preserves the real-world statistical relationship between
396 sociodemographic variables and vaccine uptake. This is demonstrated by comparing logistic
397 regression coefficients that explain the relationship between these variables and vaccine uptake
398 across the original survey populations, the synthetic population generated with our approach, and
399 the null model. As shown in Table 4, in the HPS dataset, real individuals who are white, male, or low-
400 income (less than \$25,000) have lower vaccination rates ($\beta = -0.2277, -0.0595, -0.5776$, respectively),
401 while those who are 65 or older and hold a bachelor's degree or higher ($\beta = 1.1670, 1.2429$,

402 respectively) are more likely to be vaccinated. The direction and the relative strength of these
403 associations are also reflected in the synthetic population generated using the HPS dataset. In
404 contrast, the null model fails to capture these underlying statistical relationships. For example, in the
405 synthetic population created by the null model, agents aged 65+ are less likely to be vaccinated ($\beta =$
406 -0.11). Similarly, while a bachelor's degree or higher is strongly associated with increased vaccine
407 uptake ($\beta = 1.24$) in the HPS data, the null model results in a weaker association between education
408 attainment and vaccine uptake ($\beta = 0.24$).

409 Similar results are presented with the synthetic population generated from the local survey dataset
410 and the corresponding population from the null model (Table 5). In the local survey, individuals who
411 are either aged 65 and older, have a bachelor's degree or higher, or with high income (greater than
412 \$100,000) are more likely to be vaccinated ($\beta = 1.0524, 0.6248, 0.4058$, respectively), while
413 individuals who are white are less likely to be vaccinated ($\beta = -0.1511$). These associations are
414 reflected in the synthetic population generated using our approach. However, the null model does
415 not capture the strong positive relationship between age 65+ and vaccine uptake observed in the
416 local survey dataset ($\beta = 1.0524$), with the coefficient becoming negative and close to zero ($\beta = -$
417 0.0841).

418 It is important to note that the logistic regression examples illustrate how the synthetic populations
419 generated with our public health approach are compared to the real populations from the surveys.
420 Variables for comparison were selected based on their significance in the original surveys, while
421 gender was excluded from the local survey dataset comparison due to its lack of significance at a
422 90% confidence level. While the strength and direction of the association between
423 sociodemographic variables and vaccine uptake were preserved in the synthetic populations from
424 both the HPS and local survey datasets, the coefficient of determination (R^2) for the logistic
425 regression also remained relatively consistent, indicating a similar fit between sociodemographic

426 variables and vaccine uptake. Specifically, the R^2 was 0.08 for the HPS dataset and 0.11 for the
 427 corresponding synthetic population, while it was 0.05 for the local survey dataset and the
 428 corresponding synthetic population. In contrast, the null models produced a much lower R^2 of 0.01.

429 **Table 4. Coefficients from logistic regression models describing the relationship between demographic variables**
 430 **and vaccine uptake from the HPS, along with the synthetic population and null model generated from the HPS.**
 431 **Significant coefficients are indicated with an asterisk (*) at the 90% confidence level (p-value < 0.10).**

Variable: Descriptor	HPS Data: $R^2 = 0.08$ Coefficient β	Synthetic Population: $R^2 = 0.11$ Coefficient β	Null Model: $R^2 = 0.01$ Coefficient β
Intercept	1.8615*	- 0.3245*	- 0.2365*
Gender: Male	- 0.0595*	- 0.1872*	0.0025
Race/Ethnicity: White	- 0.2277*	- 0.2758*	- 0.2761*
Age: 65 and over	1.1670*	1.4072*	- 0.1128*
Education: Bachelors or Higher	1.2429*	1.3435*	0.2357*
Income: \$25,000 or less	- 0.5776*	- 0.8446*	- 0.3071*

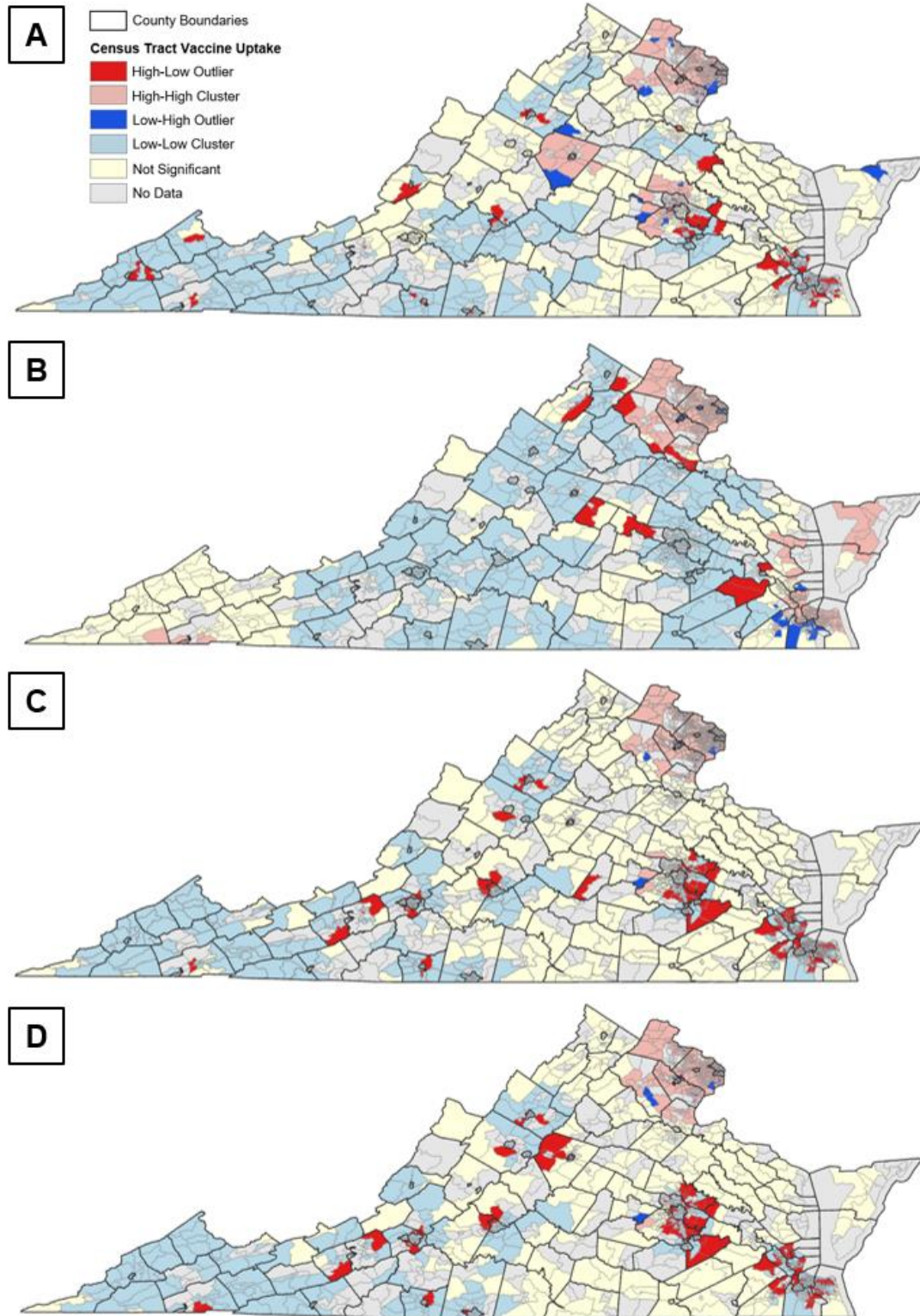
432

433 **Table 5. Coefficients from logistic regression models describing the relationship between demographic variables**
 434 **and vaccine uptake from the local survey, the synthetic population generated from the local survey, and the null**
 435 **model generated from the HPS. Significant coefficients are indicated with an asterisk (*) at the 90% confidence level**
 436 **(p-value < 0.10); Gender was not included due to its insignificance (p-value greater than 0.10) in the local survey**
 437 **dataset.**

Variable: Descriptor	Local Data: $R^2 = 0.05$ Coefficient β	Synthetic Population: $R^2 = 0.05$ Coefficient β	Null Model: $R^2 = 0.01$ Coefficient β
Intercept	0.2341*	0.4020*	- 0.4055*
Race/Ethnicity: White	- 0.1511*	- 0.3036*	- 0.2964*
Age: 65 and over	1.0524*	1.1722*	- 0.0841*
Education: Bachelors or Higher	0.6248*	0.3839*	0.1410*
Income: Greater than \$100,000	0.4058*	0.6669*	0.3703*

438

439



440

441

442 **Figure 4. Clusters and outliers of vaccination uptake in Virginia Census Tracts: A) observed, B) null model, and C)**

443 **synthesized population with HPS data, and D) synthesized population with local survey data.**

444 Figure 4 illustrates the spatial heterogeneity of vaccination uptake across the 1,592 census tracts for
445 which data was available for both real and synthetic populations. In these maps, a “High-High
446 Cluster” (light pink) indicates that census tracts have high vaccine uptake and are surrounded by
447 counties with similarly high vaccine uptake. In contrast, a “Low-Low Cluster” (light blue) represents
448 census tracts with low vaccine uptake and are surrounded by counties also with low vaccine uptake.
449 Outlier census tracts are identified as “High-Low Outliers” (bright red), where census tracts with high
450 uptake are surrounded by those with low uptake, or “Low-High Outliers” (bright blue), where census
451 tracts with low uptake are surrounded by those with high uptake. Census tracts without a significant
452 relationship to their neighbors are shown in light yellow, while those with no population or vaccine
453 data are in grey.

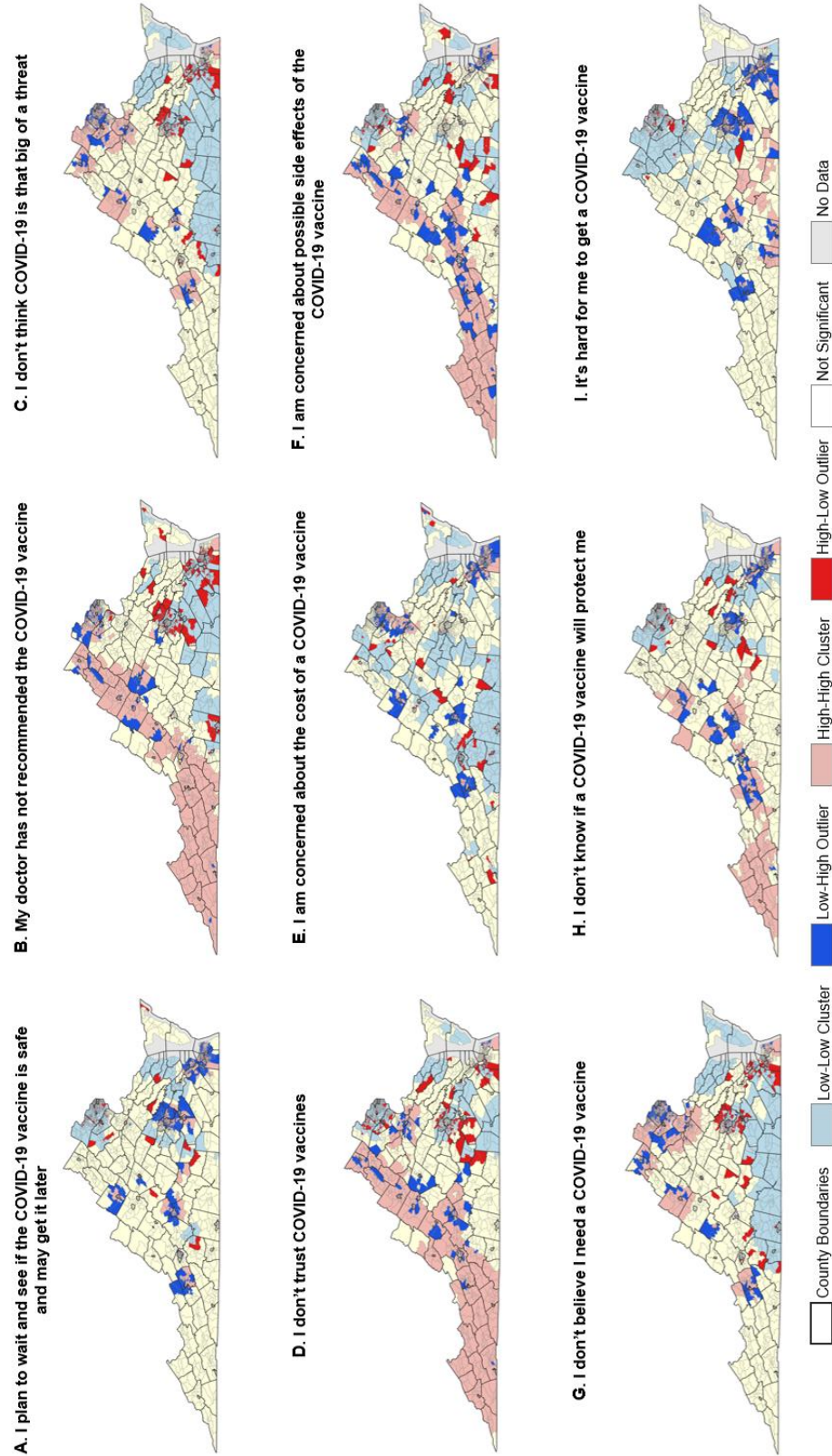
454 The observed vaccine uptake by December 2021 is mapped in Figure 4A. Generally, census tracts in
455 the western part of Virginia show relatively low vaccine uptake. Clusters of tracts with high vaccine
456 uptake are found in Northern Virginia, including Fairfax, Prince William, Loudoun, and Arlington
457 Counties. Other high uptake clusters appear in the central part of the state, such as Albemarle
458 County, which surrounds Charlottesville, and Hanover County, particularly in census tracts west of
459 Richmond. The rest of Virginia exhibits mixed uptake rates, leading to the formation of outliers. These
460 outliers are scattered throughout the state, with many High-Low outliers concentrated in larger
461 areas, such as southeast of the Richmond metropolitan area, around Hampton Roads, and in
462 smaller regions near major cities like Harrisonburg and Forest.

463 Generally, the population generated using the null model approach captures the spatial
464 heterogeneity of COVID-19 vaccine uptake since the marginal totals of the synthetic population
465 vaccination are imposed to match the real county-level data (Figure 4B). However, given that only
466 county-level data is publicly available, there is less within-county variation. For instance, the null
467 model accurately detects the high vaccine uptake cluster in Northern Virginia but inaccurately

468 suggests similar clusters in the southeastern region and census tracts along the Chesapeake Bay.
469 Additionally, the null model overlooks the low vaccine uptake clusters in southwestern Virginia. It
470 also fails to replicate the general outlier patterns observed in real vaccine uptake (Figure 4A), and
471 specifically misclassifies High-Low outliers in southeastern Virginia as Low-High outliers. This
472 indicates that imposing vaccine decisions during the initialization of agent populations does not
473 adequately preserve the spatial distribution of protective behaviors.

474 The synthetic populations generated using the HPS survey (Figure 4C) and the local survey (Figure
475 4D) generally align better with observed vaccine rates compared to the null model. They effectively
476 capture the high vaccination cluster in Northern Virginia and the low vaccine cluster in the
477 southwest. However, they fall short in replicating the larger high vaccination clusters in central
478 Virginia near Charlottesville and Richmond. Despite this, our approach excels in preserving both
479 broad regional patterns and location-specific outliers. For example, the High-Low outliers in the
480 Hampton Roads and Richmond metropolitan areas, as well as certain census tracts near
481 Harrisonburg and Forest, are accurately reflected in the synthetic populations. Notably, our method
482 also identifies the sole Low-High outlier census tract west of Richmond, an area primarily
483 characterized by High-Low outliers. These findings highlight the effectiveness of our approach in
484 capturing the spatial heterogeneity of protective behaviors.

It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).



485

486

487

Figure 5. Cluster and outliers of vaccination attitudes, beliefs, and perceptions within Virginia Census Tracts for the synthetic population generated from the HPS.

488 With our approach, all variables from the public health survey are incorporated into the agent
489 population, enabling us to generate synthetic populations with not only initial uptake of protective
490 behaviors like vaccination but also realistic attitudes, beliefs, and perceptions. This allows for a
491 better understanding of the spatial patterns of these characteristics within a population. Figure 5
492 illustrates the vaccination attitudes, beliefs, and perceptions of the synthetic population generated
493 from the HPS survey. For example, in western Virginia, clusters of individuals exhibit vaccine
494 hesitancy due to reasons such as lack of doctor recommendation (Figure 5B), distrust in the vaccine
495 (Figure 5D), or concerns about side effects (Figure 5F). In contrast, Northern Virginia shows a low
496 clustering of individuals planning to wait to see if the vaccine is safe (Figure 5A) or doubting its
497 efficacy (Figure 5H). However, there is a high concentration of individuals who do not perceive
498 COVID-19 as a significant threat (Figure 5C) or do not feel the need for the vaccine (Figure 5G).
499 Concerns about vaccine cost are prevalent in eastern Northern Virginia and extend slightly south, as
500 well as in the southeast around the Hampton Roads region (Figure 5E). Specific census tracts with
501 individuals that believe it is hard for them to get a vaccine are shown in Figure 5I. This approach
502 facilitates the integration of behavioral theories, such as the Health Belief Model (HBM), into ABMs
503 by illustrating how individual attitudes, beliefs, and perceptions affect vaccine uptake and spatial
504 distribution. This capability ultimately supports the development of ABMs of infectious disease
505 spread that aim to simulate the underlying processes driving the adoption of protective behaviors
506 over time, providing a realistic initialization of populations with these characteristics.

507 **5. Discussion and Conclusion**

508 In this study, we investigate the potential to expand synthetic population generation approaches to
509 initialize an agent population with variables relevant for public health, using COVID-19 vaccine
510 uptake as an example. Our results show that such an approach has potential to support disease
511 simulations requiring realistic parameterization of agents with these variables. This method enables

512 researchers to quickly initialize a synthetic population where the true statistical relationships
513 between demographic characteristics and public health variables are preserved. Furthermore, the
514 approach captures the spatial heterogeneity of such protective behaviors at finer scales than
515 typically available in spatial data. While protective behaviors such as vaccination, masking, and
516 social distancing can sometimes be found at county or state level, similar data capturing attitudes,
517 beliefs and perceptions that can be simulated using this approach are often not available in spatial
518 data format at all. The synthetic population that was generated using the national HPS that is publicly
519 available was comparable to the synthetic population generated using the local survey data,
520 demonstrating the flexibility of the approach to be implemented using a variety of public health
521 surveys.

522 It is important to note that researchers who have access to fine-grained spatial data capturing health
523 behavior variables (e.g. vaccine uptake at the CT level) could incorporate that data directly into the
524 IPF approach to more accurately capture the statistical and spatial patterns health behaviors.
525 However, this would be limited to the study area and time for which the data is available. For example,
526 our validation dataset captures vaccine uptake at the CT level for Virginia by December 2021,
527 meaning it could be used in the IPF as another category for which the marginal totals are known.
528 However, this would limit the transferability of the approach to other study areas and points in time.
529 Therefore, we demonstrate how the approach could be implemented using only publicly available
530 longitudinal data such as the HPS, making it straightforward for researchers to generate a synthetic
531 population with these variables anywhere in the country and for multiple points in time.

532 The quality of the synthetic population is limited by the quality of the census tract and individual level
533 survey data. For example, it does not appear that the HPS data is nationally representative and was
534 largely biased towards vaccinated individuals. Therefore, we were able to improve our results slightly
535 using this dataset by adjusting the representation of vaccinated individuals in the sample from 91%

536 ($r^2 = 0.499$) to 50% ($r^2 = 0.56$), but more research is needed to investigate effects of bias on this
537 approach. Furthermore, there are other factors that are likely affecting the individual's decision to
538 get vaccinated (e.g. policy interventions, social norms) that can't be directly incorporated into the
539 synthetic population generation approach. Thus, it may be more effective to synthesize a population
540 with vaccine intention, rather than vaccine uptake itself, where such data is available (e.g. using the
541 Understanding Coronavirus in America longitudinal survey).

542 In conclusion, this study demonstrates the potential of using public health surveys to enhance the
543 generation of synthetic populations for spatial agent-based models (ABMs) by incorporating
544 protective behaviors and attitudes. Such an approach strengthens the potential for ABM in public
545 health research and policy planning. We show that the synthetic populations generated using this
546 approach reflect the real-world statistical relationships between demographic groups and vaccine
547 uptake and attitudes. This level of detail is essential for simulating potential health disparities across
548 different demographic groups, enabling the exploration of more targeted and effective public health
549 strategies. Initializing an ABM with realistic vaccine uptake and attitudes is crucial for those that aim
550 to forecast outbreaks in study areas where uptake and attitudes vary regionally or that aim to
551 simulate realistic social processes driving vaccine uptake decisions. By capturing the spatial
552 heterogeneity of these behaviors and attitudes at finer scales than spatial data typically allows, our
553 approach supports models that aim to simulate how local responses to interventions might unfold
554 with greater accuracy and how these responses lead to spatially heterogeneous health outcomes.
555 Ultimately, this approach enhances the predictive power and realism of disease simulations,
556 providing critical insights into how interventions might play out in real-world settings.

557 To our knowledge, this study marks one of the first attempts to extend synthetic population
558 generation approaches to initialize agents with variables relevant to ABMs of infectious disease
559 spread. Future research is needed to see if this approach can be used to initialize other health

560 behaviors and associated perceptions and attitudes (e.g., tobacco use in populations using the
561 CDC's National Tobacco Survey). We encourage other researchers with access to more fine-grained
562 spatially aggregated data to validate this approach across various public health domains, to improve
563 the parameterization of more realistic agent populations in data-driven ABMs for public health.

564 **Acknowledgements**

565 This research was funded by National Science Foundation (Award #230970 and #2109647).

566 **Data availability statement**

567 The data including the validation dataset and code for generating the synthetic population based on
568 the Household Pulse Survey is available on a GitHub repository at
569 <https://github.com/evonhoene/Population-Generation-for-Public-Health-ABMs>. The spatially
570 aggregate data from the American Community Survey and the individual-level Household Pulse
571 Survey are publicly available and the links are provided on the GitHub page. De-identified data from
572 the local Virginia survey data is available upon request.

573 **References**

- 574 1. Pesavento J, Chen A, Yu R, Kim JS, Kavak H, Anderson T, et al. Data-driven mobility models for
575 COVID-19 simulation. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on
576 Advances in Resilient and Intelligent Cities [Internet]. Seattle Washington: ACM; 2020 [cited
577 2023 Jun 27]. p. 29–38. Available from: <https://dl.acm.org/doi/10.1145/3423455.3430305>
- 578 2. Von Hoene E, Roess A, Achuthan S, Anderson T. A Framework for Simulating Emergent Health
579 Behaviors in Spatial Agent-Based Models of Disease Spread. In: Proceedings of the 6th ACM
580 SIGSPATIAL International Workshop on GeoSpatial Simulation [Internet]. Hamburg Germany:
581 ACM; 2023 [cited 2024 Apr 16]. p. 1–9. Available from:
582 <https://dl.acm.org/doi/10.1145/3615891.3628010>
- 583 3. Bicher M, Rippinger C, Urach C, Brunmeir D, Siebert U, Popper N. Evaluation of Contact-Tracing
584 Policies against the Spread of SARS-CoV-2 in Austria: An Agent-Based Simulation. *Med Decis*
585 *Making*. 2021 Nov 1;41(8):1017–32.
- 586 4. Hunter E, Kelleher JD. Adapting an Agent-Based Model of Infectious Disease Spread in an Irish
587 County to COVID-19. *Systems*. 2021 Jun;9(2):41.

- 588 5. Anderson T, Dragičević S. NEAT approach for testing and validation of geospatial network agent-
589 based model processes: case study of influenza spread. *Int J Geogr Inf Sci.* 2020;34(9):1792–
590 821.
- 591 6. Adiga A, Chu S, Eubank S, Kuhlman CJ, Lewis B, Marathe A, et al. Disparities in spread and
592 control of influenza in slums of Delhi: findings from an agent-based modelling study. *BMJ Open.*
593 2018 Jan 21;8(1):e017353.
- 594 7. Rafferty ERS, McDonald W, Osgood ND, Qian W, Doroshenko A. Seeking the optimal schedule
595 for chickenpox vaccination in Canada: Using an agent-based model to explore the impact of
596 dose timing, coverage and waning of immunity on disease outcomes. *Vaccine.* 2020 Jan
597 16;38(3):521–9.
- 598 8. Tang X, Zhao S, Chiu APY, Ma H, Xie X, Mei S, et al. Modelling the transmission and control
599 strategies of varicella among school children in Shenzhen, China. *PLOS ONE.* 2017 May
600 18;12(5):e0177514.
- 601 9. Bian L. A Conceptual Framework for an Individual-Based Spatially Explicit Epidemiological
602 Model. *Environ Plan B Plan Des.* 2004 Jun 1;31(3):381–95.
- 603 10. Heppenstall AJ, Crooks AT, See LM, Batty M, editors. *Agent-Based Models of Geographical
604 Systems* [Internet]. Dordrecht: Springer Netherlands; 2012 [cited 2024 Aug 10]. Available from:
605 <https://link.springer.com/10.1007/978-90-481-8927-4>
- 606 11. Buckee C, Noor A, Sattenspiel L. Thinking clearly about social aspects of infectious disease
607 transmission. *Nature.* 2021 Jul;595(7866):205–13.
- 608 12. Eisenstein M. Disease: Poverty and pathogens. *Nature.* 2016 Mar;531(7594):S61–3.
- 609 13. House T, Keeling MJ. Household structure and infectious disease transmission. *Epidemiol
610 Infect.* 2009 May;137(5):654–61.
- 611 14. Geard N, McCaw JM, Dorin A, Korb KB, McVernon J. Synthetic Population Dynamics: A Model of
612 Household Demography. *J Artif Soc Soc Simul.* 2013;16(1):8.
- 613 15. Duerr HP, Schwehm M, Leary CC, Vlas SJD, Eichner M. The impact of contact structure on
614 infectious disease control: influenza and antiviral agents. *Epidemiol Infect.* 2007
615 Oct;135(7):1124–32.
- 616 16. Zhu K, Yin L, Liu K, Liu J, Shi Y, Li X, et al. Generating synthetic population for simulating the
617 spatiotemporal dynamics of epidemics. *PLOS Comput Biol.* 2024 Feb 12;20(2):e1011810.
- 618 17. Del Valle SY, Hyman JM, Hethcote HW, Eubank SG. Mixing patterns between age groups in
619 social networks. *Soc Netw.* 2007 Oct 1;29(4):539–54.
- 620 18. Lovelace R, Birkin M, Ballas D, van Leeuwen E. Evaluating the Performance of Iterative
621 Proportional Fitting for Spatial Microsimulation: New Tests for an Established Technique. *J Artif
622 Soc Soc Simul.* 2015;18(2):21.

- 623 19. National Household Travel Survey [Internet]. US Department of Transportation Federal Highway
624 Administration; [cited 2024 Jul 2]. Available from: <https://nhts.ornl.gov/>
- 625 20. Bureau UC. Census.gov. [cited 2023 Jul 12]. Public Use Microdata Sample (PUMS). Available
626 from: <https://www.census.gov/programs-surveys/acs/microdata.html>
- 627 21. d’Andrea V, Gallotti R, Castaldo N, Domenico MD. Individual risk perception and empirical
628 social structures shape the dynamics of infectious disease outbreaks. *PLOS Comput Biol*. 2022
629 Feb 16;18(2):e1009760.
- 630 22. Funk S, Salathé M, Jansen VAA. Modelling the influence of human behaviour on the spread of
631 infectious diseases: a review. *J R Soc Interface*. 2010 May 26;7(50):1247–56.
- 632 23. Funk S, Bansal S, Bauch CT, Eames KTD, Edmunds WJ, Galvani AP, et al. Nine challenges in
633 incorporating the dynamics of behaviour in infectious diseases models. *Epidemics*. 2015 Mar
634 1;10:21–5.
- 635 24. Manfredi P, D’Onofrio A. Modeling the Interplay Between Human Behavior and the Spread of
636 Infectious Diseases. Springer Science & Business Media; 2013. 329 p.
- 637 25. de Mooij J, Bhattacharya P, Dell’Anna D, Dastani M, Logan B, Swarup S. A framework for
638 modeling human behavior in large-scale agent-based epidemic simulations. *Simul-Trans Soc
639 Model Simul Int*. 2023 Dec;99(12):1183–211.
- 640 26. Retzlaff CO, Burbach L, Kojan L, Halbach P, Nakayama J, Ziefle M, et al. Fear, Behavior, and the
641 COVID-19 Pandemic: A City-Scale Agent-Based Model Using Socio-Demographic and Spatial
642 Map Data. *JASSS- J Artif Soc Soc Simul*. 2022 Jan;25(1).
- 643 27. Naugle A, Rothganger F, Verzi S, Doyle C. Conflicting Information and Compliance with COVID-
644 19 Behavioral Recommendations. *JASSS- J Artif Soc Soc Simul*. 2022 Oct;25(4).
- 645 28. Alvarez-Zuzek LG, Rocca CEL, Iglesias JR, Braunstein LA. Epidemic spreading in multiplex
646 networks influenced by opinion exchanges on vaccination. *PLOS ONE*. 2017 Nov
647 9;12(11):e0186492.
- 648 29. Pandey A, Fitzpatrick MC, Moghadas SM, Vilches TN, Ko C, Vasan A, et al. Modelling the impact
649 of a high-uptake bivalent booster scenario on the COVID-19 burden and healthcare costs in
650 New York City. *Lancet Reg Health – Am* [Internet]. 2023 Aug 1 [cited 2024 Aug 8];24. Available
651 from: [https://www.thelancet.com/journals/lanam/article/PIIS2667-193X\(23\)00129-1/fulltext](https://www.thelancet.com/journals/lanam/article/PIIS2667-193X(23)00129-1/fulltext)
- 652 30. Crooks AT, Heppenstall AJ. Introduction to Agent-Based Modelling. In: Heppenstall AJ, Crooks
653 AT, See LM, Batty M, editors. *Agent-Based Models of Geographical Systems* [Internet].
654 Dordrecht: Springer Netherlands; 2012 [cited 2024 Aug 10]. p. 85–105. Available from:
655 https://doi.org/10.1007/978-90-481-8927-4_5
- 656 31. Chapuis K, Taillandier P, Drogoul A. Generation of Synthetic Populations in Social Simulations:
657 A Review of Methods and Practices. *J Artif Soc Soc Simul*. 2022;25(2):6.

- 658 32. Yameogo BF, Vandanjon PO, Gastineau P, Hankach P. Generating a Two-Layered Synthetic
659 Population for French Municipalities: Results and Evaluation of Four Synthetic Reconstruction
660 Methods. *J Artif Soc Soc Simul*. 2021;24(2):5.
- 661 33. Kotnana S, Han D, Anderson T, Züfle A, Kavak H. Using Generative Adversarial Networks to
662 Assist Synthetic Population Creation for Simulations. In: 2022 Annual Modeling and Simulation
663 Conference (ANNSIM) [Internet]. 2022 [cited 2024 Aug 23]. p. 1–12. Available from:
664 <https://ieeexplore.ieee.org/document/9859422>
- 665 34. Huang Z, Williamson P. A comparison of synthetic reconstruction and combinatorial
666 optimisation approaches to the creation of small-area microdata. *Dep Geogr Univ Liverp*. 2001;
- 667 35. Ye X, Konduri K, Pendyala R, Sana B, Waddell P. Methodology to match distributions of both
668 household and person attributes in generation of synthetic populations. 2009 Jan 1;
- 669 36. Pritchard DR, Miller EJ. Advances in population synthesis: fitting many attributes per agent and
670 fitting to household and person margins simultaneously. *Transportation*. 2012 May 1;39(3):685–
671 704.
- 672 37. Deming WE, Stephan FF. On a Least Squares Adjustment of a Sampled Frequency Table When
673 the Expected Marginal Totals are Known. *Ann Math Stat*. 1940;11(4):427–44.
- 674 38. Beckman RJ, Baggerly KA, McKay MD. Creating synthetic baseline populations. *Transp Res Part
675 Policy Pract*. 1996 Nov 1;30(6):415–29.
- 676 39. Lovelace R, Ballas D. ‘Truncate, replicate, sample’: A method for creating integer weights for
677 spatial microsimulation. *Comput Environ Urban Syst*. 2013 Sep 1;41:1–11.
- 678 40. Moradi S. RecovUS: An agent-based model of post-disaster housing recovery. 2020 May [cited
679 2024 Aug 14]; Available from: <https://hdl.handle.net/2346/85845>
- 680 41. Zhu K, Yin L, Liu K, Liu J, Shi Y, Li X, et al. Generating synthetic population for simulating the
681 spatiotemporal dynamics of epidemics. *PLOS Comput Biol*. 2024 Feb 12;20(2):e1011810.
- 682 42. Alagoz O, Sethi AK, Patterson BW, Churpek M, Safdar N. Effect of Timing of and Adherence to
683 Social Distancing Measures on COVID-19 Burden in the United States. *Ann Intern Med*. 2020
684 Oct 27;M20-4096.
- 685 43. Eilersen A, Sneppen K. Cost–benefit of limited isolation and testing in COVID-19 mitigation |
686 Scientific Reports. *Sci Rep* [Internet]. 2020 [cited 2024 Aug 15];10. Available from:
687 <https://www.nature.com/articles/s41598-020-75640-2>
- 688 44. Luo W, Gao P, Cassels S. A large-scale location-based social network to understanding the
689 impact of human geo-social interaction patterns on vaccination strategies in an urbanized
690 area. *Comput Environ Urban Syst*. 2018 Nov 1;72:78–87.
- 691 45. Von Hoene E, Roess A, Achuthan S, Anderson T. A Framework for Simulating Emergent Health
692 Behaviors in Spatial Agent-Based Models of Disease Spread. In: Proceedings of the 6th ACM
693 SIGSPATIAL International Workshop on GeoSpatial Simulation [Internet]. New York, NY, USA:

- 694 Association for Computing Machinery; 2023 [cited 2024 Jul 4]. p. 1–9. (GeoSim '23). Available
695 from: <https://doi.org/10.1145/3615891.3628010>
- 696 46. Mao L. Predicting Self-Initiated Preventive Behavior Against Epidemics with an Agent-Based
697 Relative Agreement Model. *J Artif Soc Soc Simul*. 2015;18(4):6.
- 698 47. Stapelberg NJC, Smoll NR, Randall M, Palipana D, Bui B, Macartney K, et al. A Discrete-Event,
699 Simulated Social Agent-Based Network Transmission (DESSABNeT) model for communicable
700 diseases: Method and validation using SARS-CoV-2 data in three large Australian cities. *PLOS*
701 *ONE*. 2021 May 21;16(5):e0251737.
- 702 48. Sinclair DR, Grefenstette JJ, Krauland MG, Galloway DD, Frankeny RJ, Travis C, et al. Forecasted
703 Size of Measles Outbreaks Associated With Vaccination Exemptions for Schoolchildren. *JAMA*
704 *Netw Open*. 2019 Aug 21;2(8):e199768.
- 705 49. Tomizawa N, Kumamaru KK, Okamoto K, Aoki S. Multi-agent system collision model to predict
706 the transmission of seasonal influenza in Tokyo from 2014–2015 to 2018–2019 seasons.
707 *Heliyon* [Internet]. 2021 Aug 1 [cited 2024 Aug 15];7(8). Available from:
708 [https://www.cell.com/heliyon/abstract/S2405-8440\(21\)01962-9](https://www.cell.com/heliyon/abstract/S2405-8440(21)01962-9)
- 709 50. American Community Survey (ACS) [Internet]. US Census Bureau; [cited 2023 Jul 6]. Available
710 from: <https://www.census.gov/programs-surveys/acs>
- 711 51. AlShurman BA, Khan AF, Mac C, Majeed M, Butt ZA. What Demographic, Social, and Contextual
712 Factors Influence the Intention to Use COVID-19 Vaccines: A Scoping Review. *Int J Environ Res*
713 *Public Health*. 2021 Sep 4;18(17):9342.
- 714 52. Household Pulse Survey [Internet]. US Census Bureau; [cited 2024 Aug 8]. Available from:
715 <https://www.census.gov/householdpulsedata>
- 716 53. COVID-19 Vaccinations in the United States,County | Data | Centers for Disease Control and
717 Prevention [Internet]. [cited 2024 Aug 8]. Available from:
718 [https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-](https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh/about_data)
719 [amqh/about_data](https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh/about_data)
- 720