Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

EVALUATING THE TRANSFERABILITY OF A SYNTHETIC POPULATION GENERATION APPROACH FOR PUBLIC HEALTH APPLICATIONS

Emma Von Hoene[1], Aanya Gupta[2], Hamdi Kavak[3], Amira Roess[4], and Taylor Anderson[1]

[1]Dept. of Geography and Geoinformation Science, George Mason University, Fairfax, USA
[2]Thomas Jefferson High School for Science and Technology, Alexandria, USA
[3]Dept. of Computational and Data Sciences, George Mason University, Fairfax, USA
[4]Dept. of Global and Community Health, George Mason University, Fairfax, USA

ABSTRACT

Simulations are valuable in public health research, with synthetic populations enabling realistic policy analysis. However, methods for generating synthetic populations with domain-specific characteristics remain underexplored. To address this, we previously introduced a population synthesis approach that directly integrates health surveys. This study evaluates its transferability across health outcomes, locations, and timeframes through three case studies. The first generates a Virginia population (2021) with COVID-19 vaccine intention, comparing results to probabilistic and regression-based approaches. The second synthesizes populations with depression (2021) for Virginia, Tennessee, and New Jersey. The third constructs Virginia populations with smoking behaviors for 2021 and 2022. Results demonstrate the method's transferability for various health applications, with validation confirming its ability to capture accuracy, statistical relationships, and spatial heterogeneity. These findings enhance population synthesis for public health simulations and offer new datasets with small-area estimates for health outcomes, ultimately supporting public health decision-making.

1 INTRODUCTION

Accurately simulating human behavior in large-scale social simulations, such as agent-based models (ABMs) or microsimulations, relies on capturing fine-grained heterogeneity within the simulated population (Axelrod 2007). Synthetic population generation is the process of creating detailed and representative artificial datasets of human individuals or households (Chapuis et al. 2022). These artificial populations reflect statistical properties of real-world demographic and geographic distributions while preserving individual-level anonymity. Synthetic populations are important inputs for models applied to domains where the outcomes are highly dependent on the heterogeneous characteristics of the individuals in the population. For example, demographic variables such as occupation, education, gender, marital status, race, and age are often determinants for travel behaviors (Bigi et al. 2024), social contact networks (Prem et al. 2017), and chronic or infectious disease prevalence (Zhu et al. 2024). Thus, failure to capture population heterogeneity could limit the practical use of the simulation results (Bigi et al. 2024; Zhu et al. 2024).

There are many approaches to population synthesis, with a comprehensive review of these methods provided by Chapuis et al. (2022). This study uses iterative proportional fitting (IPF), a well-established and widely used technique (Huang and Williamson 2001). IPF typically requires two inputs: a disaggregated sample of the population with demographic characteristics and spatially aggregated data that captures the demographic marginal totals for each spatial zone (e.g., census block group, census tract). Using this data, IPF calculates a weight for each individual in the sample based on how well their demographic characteristics align with the spatially aggregated constraints. These weights are then converted into integers using an 'integerisation' method (Lovelace and Ballas 2013), which determines how many times each individual should be replicated within each census tract. This is followed by expansion, in which individuals are

duplicated according to their integer weights and assigned to the appropriate geographic zones. Lastly, in the replication stage, the set of attributes from the original individual-level dataset is joined to each synthesized individual. The resulting synthetic population includes individuals linked to a spatial zone and characterized by demographic variables such as age, gender, race, education and income.

While most population synthesis efforts focus on specifying demographic attributes, simulations applied to public health often require a synthetic population enriched with additional variables such as underlying health conditions or health perceptions and attitudes. For instance, a model aiming to simulate cardiovascular disease needs additional attributes that contribute to the disease including smoking, diabetes, and medication status (Knight et al. 2017). Generating a realistic synthetic population with these characteristics can offer insights for addressing health disparities and supporting public health interventions.

Techniques aiming to enrich synthetic populations for public health applications commonly use probabilistic or regression-based approaches to assign health-relevant attributes after generating a base population with only demographic characteristics. For example, Pandey et al. (2023) assign COVID-19 vaccine coverage to a New York City population based on an agent's age using probabilities derived from historical influenza data. Similarly, Nicolaie et al. (2023) apply a series of regression models to sequentially predict new variables, using previously predicted features, resulting in a Dutch population with demographic and lifestyle variables as well as disease prevalence. While these methods offer flexibility, they can be cumbersome to implement, suffer from cascading errors, and may fail to maintain joint distributions across multiple variables. For instance, Nicolaie et al. (2023) found that while the features predicted in the early stages of their sequential procedure were highly accurate, accuracy decreased in later stages due to the accumulation of errors and uncertainty in previously modeled features.

A streamlined alternative is to directly integrate public health survey data into the population synthesis process, which we refer to as a *direct survey-based approach*. This approach supports the goal of synthetic population generation by creating realistic populations that jointly capture demographic and health attributes along with their geographic distribution, while addressing limitations of commonly used data sources. Census data provide detailed spatial coverage but lack health attributes, whereas health surveys offer rich health information but are disaggregated and lack spatial detail. For example, Von Hoene et al. (2025) proposed using COVID-19 vaccine-related surveys, instead of typically used census-based disaggregated demographic data, to initialize agents with vaccine decisions and attitudes in Virginia, U.S. This method synthesized populations that reflected real-world spatial variation and the underlying associations between demographics and vaccine uptake. However, the performance of this approach compared to traditional methods, as well as its generalizability to other health outcomes, time periods, or locations, remains unexplored.

To explore this further, we developed a series of synthetic populations (Section 2) to evaluate the direct survey-based approach presented in Von Hoene et al. (2025). This study first compares the performance of the direct survey-based approach against two baseline methods, including a regression-based and a probabilistic approach, using vaccine behaviors as a case study. Next, it examines the approach's transferability for initializing populations with other public health attributes, specifically smoking and depression, across different locations and time periods. The results (Section 3) demonstrate how the synthesized populations capture overall accuracy, preserve statistical relationships with demographic attributes, and reflect true spatial heterogeneity. As discussed in Section 4, we argue that the direct survey-based approach is an advantageous method for supporting social simulations and decision-making as it allows researchers to effectively generate synthetic populations across various health outcomes, locations, and time periods.

2 METHODOLOGY

We evaluate the transferability of the direct-survey-based approach through three experiments: 1) comparing it with two baseline approaches, including a regression-based and a probabilistic approach, using vaccine behaviors as a case study; 2) adapting the approach for other public health applications and study areas, using the case study of clinical depression; and 3) extending the approach to multiple research purposes and time points, using the case study of cigarette smoking. For each of these experiments, we generate

the synthetic populations with the approach proposed by Von Hoene et al. (2025) that uses IPF; refer to the study for further details on the method. This section describes the data and experiments, as follows.

2.1 Data

2.1.1 Input Data

Recall that IPF-based method in Von Hoene at al. (2025) relies on both individual-level survey data as a disaggregated sample of the population and spatially aggregated demographic data capturing marginal population totals. The datasets used for the generation of the synthetic populations are described as follows:

Individual level survey data. This study relies on nationally representative survey data from the Understanding America Study (UAS), administered by the Center for Economic and Social Research at the University of Southern California (University of Southern California 2025). The UAS collects data on various topics from a panel of individuals aged 18 and older, with most data publicly available.

Given our three case studies—vaccine intention, depression, and smoking—we use data from the following UAS surveys that capture each attribute: (1) the Understanding Coronavirus in America (UAS-UCA), which collects data about individuals during the COVID-19 pandemic in the U.S., and (2) the Drug Use Supplement (UAS-DUS), which tracks monthly variations in drug use patterns and related characteristics. For the UAS-UCA, we specifically use UAS 282 – Wave 24, in which 6,344 individuals participated between February 3 and March 2, 2021. For UAS-DUS, we use UAS 423 – Wave 36, in which 7,165 individuals participated between February 21 and March 20, 2022. The UAS-UCA survey was used to generate synthetic populations with COVID-19 vaccine intention (case study 1), diagnosed depression (case study 2), and smoking behaviors (case study 3) representative of March 2021. In this survey, 66.9% of respondents reported an intent to receive the COVID-19 vaccine, 21.4% reported a diagnosis of depression, and 11.6% identified as current cigarette smokers. Additionally, the UAS-DUS was used to synthesize a population with smoking behaviors as of March 2022, allowing us to examine changes in behavior over time. In this survey, 11.0% of respondents identified as current cigarette smokers. Any survey records with missing responses for required variables for the population synthesis were excluded, resulting in a final sample of N = 4,508 respondents for UAS-UCA and N = 5,634 respondents for UAS-DUS.

Spatially aggregated data. Spatial demographic data was sourced from the U.S. Census Bureau's American Community Survey (ACS) (U.S. Census Bureau 2025), which provides 2017-2021 5-year estimates for demographic characteristics at the census tract level. Attributes considered included age, gender, education, income, and race/ethnicity, as these characteristics are known to influence the health outcomes—specifically COVID-19 vaccine intention, smoking, and depression—synthesized in this study (AlShurman et al. 2021; Garrett et al. 2019; Akhtar-Danesh and Landeen 2007). The ACS 2017-2021 estimates for census tracts in Virginia (for all case studies), New Jersey (for case study 2), and Tennessee (for case study 2) were used. Records with missing data for any demographic variable or census tracts with a population of zero were excluded, resulting in spatially aggregated datasets with N = 2,162, N = 2,165, and N = 1,680 census tracts for Virginia, New Jersey and Tennessee, respectively.

2.1.2 Data Pre-processing

In the population synthesis method, IPF fits survey data with spatially aggregated data, requiring both datasets to share the same demographic categories. The spatially aggregated data was processed to include marginal totals for these categories: *age*, with groups 18–29, 30–49, 50–64, and 65+; *gender*, divided into male and female; *education*, categorized as no bachelor's degree or bachelor's degree and higher; *income*, with ranges including less than $25,000, $25,000–49,999, $50,000–99,999, and greater than $100,000; and *race/ethnicity*, classified as White, Black, Hispanic, and other. The survey datasets were processed to match these categories and converted into a binary format for consistency in the population synthesis approach. For example, in the agent populations, a value of "0" for male indicates the individual is not male, while a value of "1" for age 65+ indicates the individual falls within that age group.

2.1.3 Validation Data

Once the agent populations are generated—where each agent represents an individual aged 18+ in the real population and is assigned to a census tract—the dataset is aggregated back to the census tract level for validation. For example, synthetic populations can be grouped by census tract to compute the percentage of synthetic individuals diagnosed with depression, which can then be compared to an external dataset measuring depression prevalence at the census tract level. Since this study synthesizes populations with attributes related to COVID-19 vaccine intention, depression, and smoking, census tract data for these outcomes are used for validation. COVID-19 vaccine intention in the synthetic population is validated against observed vaccine uptake data for Virginia as of December 30, 2021, obtained by request from the Virginia Department of Public Health. Depression and smoking prevalence in the synthetic population are validated using the 2024 release of the CDC's PLACES dataset (Centers for Disease Control and Prevention 2025). The PLACES dataset provides model-based estimates of health-related measures based on the 2021 or 2022 Behavioral Risk Factor Surveillance System (BRFSS) data (Centers for Disease Control and Prevention 2024) and ACS 2018–2022 estimates (U.S. Census Bureau 2025), including the prevalence of diagnosed depression and current cigarette smoking among adults, for the entire United States at various geographic scales. Although PLACES data are derived from small-area estimation methods, they have been validated through internal and external studies (Zhang et al. 2015), establishing them as a reliable source for validating the synthetic populations presented in this study.

2.2 Experiments and Case Studies

2.2.1 Case Study 1: Comparing Approaches with COVID-19 Vaccine Intention

The purpose of the first experiment is to compare the direct survey-based approach (Von Hoene et al. 2025) with two other commonly used approaches, specifically regression and probabilistic methods, using vaccine intention as a case study. This case study generates three synthetic populations representing individuals aged 18 and older in Virginia (N = 6,672,836) in 2021, using the UAS-UCA from March 2021 and Virginia census tract spatially aggregated data. Each population incorporates demographic characteristics and COVID-19 vaccine intention, where each is generated using one of the following methods: the direct survey-based approach, a regression approach, or a probabilistic approach.

The *direct survey-based approach* presented in Von Hoene et al. (2025) replaces traditional datasets used in IPF, such as the Public Use Microdata Survey (PUMS), with public health surveys. This provides a straightforward method for directly replicating individuals with public health attributes based on any given survey question. In this case study, the UAS-UCA question, *"How likely are you to get vaccinated for coronavirus once a vaccine is available to the public?"*, was used to assign a COVID-19 vaccine intention variable to agents in the synthetic population generated from the direct survey-based approach.

Regression-based and probabilistic approaches are commonly used to assign public health attributes after generating a base population with demographic characteristics. Here, we describe two baseline models for comparison with our direct survey-based approach. First, we generate a base population with age, income, education, gender, and race/ethnicity using IPF. For the *baseline regression approach*, we fit a logistic regression model using the UAS-UCA input survey dataset, with demographic variables as independent variables and COVID-19 vaccine intention as the dependent variable. This regression model is then applied to each agent, using their own demographic characteristics to predict their intention to vaccinate. Given that logistic regression predicts probabilities, a threshold based on the median probability (0.6) is used to classify whether an agent intends to receive the COVID-19 vaccine.

For the *baseline probabilistic approach*, the percentage of UAS-UCA survey respondents who intended to receive the COVID-19 vaccine was calculated for different demographic combinations. These percentages were then imposed on the base population based on matching demographics. For example, 53% of survey respondents who were not male, not Black, not aged 65 or older, did not hold a bachelor's degree, and does not have an income greater than $100,000 reported an intention to get vaccinated. Accordingly,

a random 53% of agents with these demographic characteristics were assigned with COVID-19 vaccine intent. The probabilistic approach was tested in five iterations, each sequentially incorporating an additional demographic variable based on its strength of association with vaccine intention. In the first iteration, vaccine intention was assigned based only on education. In the second, both education and age were considered. This process continued until the fifth and final iteration, where vaccine intention probabilities were calculated and assigned based on the combination of education, age, income, race, and gender. Only the results from this final iteration, which incorporates all demographics, are presented in Section 3.

The generated populations were each validated using real vaccine uptake data from December 2021 across Virginia census tracts. While agents are initialized with COVID-19 vaccine intent rather than actual vaccine uptake, we make this choice because intent is likely more strongly associated with demographic characteristics, whereas actual uptake was heavily influenced by government interventions and mandates for school, work, and other activities. Vaccine uptake data was available for 1,601 census tracts in which we generated a population. Nine records with vaccine uptake greater than 100% were removed as outliers. As a result, our validation was conducted only for census tracts where data was available and vaccine uptake was 100% or less (N=1,592).

2.2.2 Case Study 2: Demonstrating Spatial Adaptability with Depression

The second experiment in this study explores whether the direct survey-based approach can be extended to other public health applications and other study areas, using the case study of clinical depression. Using the UAS-UCA survey and spatially aggregated census tract data for Virginia, New Jersey, and Tennessee, three agent populations were created, each representing individuals aged 18 and older with demographic characteristics and diagnosed depression. These populations, representative of 2021, consisted of 6,672,836 individuals in Virginia, 6,919,601 in New Jersey, and 5,257,469 in Tennessee. These states were selected based on their varying overall depression rates in the contiguous US: New Jersey has the lowest average rate, Virginia aligns with the national average, and Tennessee has the highest mean rate, according to the CDC's PLACES dataset. The synthetic populations were assigned depression attributes based on responses to the UAS-UCA survey question: *"Has a doctor or another health professional ever told you that you have depression or another depressive disorder?"* The synthesized populations for Virginia, New Jersey, and Tennessee were validated using the CDC's PLACES estimates for depression prevalence across census tracts. Specifically, the PLACES dataset provides the percentage of the population who have ever been diagnosed with a depressive disorder by a health professional, which aligns with the survey question used in the UAS-UCA to generate the synthetic populations.

2.2.3 Case Study 3: Modeling Temporal Changes with Smoking

In the third experiment, we investigate the degree to which the direct survey-based approach can be used to synthesize changes in populations at multiple points in time using smoking as a case study. Two populations representative of Virginia (N = 6,672,836) for 2021 and 2022 were created, each consisting of individuals aged 18 and older with demographic characteristics and cigarette smoking behavior based on the two UAS surveys. Both surveys included the same question: *"Out of the past seven days, what is your best estimate of the number of days that you smoked all or part of a cigarette?"* Individuals who reported smoking at least once in the past week were classified as current smokers in the synthetic populations. Each synthetic population was validated using the CDC's PLACES estimates for the number of individuals who smoke every day or some days, which closely corresponds with the smoking attribute in the generated populations.

3 RESULTS

The results of the synthetic populations generated for each case study are evaluated along three dimensions. First, we evaluate the *overall accuracy* of our simulated health outcomes—COVID-19 vaccine intent, depression, and smoking—by aggregating them to the census tract level and comparing them to the

corresponding validation datasets using Pearson's correlation coefficient (r), coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE) (Table 1). Since IPF is well known for its ability to accurately capture demographic variables, we do not report these measures for gender, race, age, education, and income, as r and R^2 values are near 1, with minimal RMSE and MAE. Second, to assess whether the synthetic populations *preserve real-world statistical relationships* between demographics and health outcomes, we compare logistic regression coefficients from the synthesized populations to those from the original survey data (Table 2). Lastly, we validate the *spatial heterogeneity* of synthesized health outcomes using spatial autocorrelation metrics, specifically the Anselin Local Moran's I statistic (Anselin 1995). This analysis identifies clusters and outliers based on the value and statistical significance ($p < 0.05$) of the I statistic, as shown in maps (Figures 1 - 3).

Table 1: Metrics comparing health outcomes in synthesized populations to validation data. Note that case studies 2 and 3 use only the direct survey-based approach to synthesize populations.

Synthesized Population: Location - Year	**Pearson's r**	**R^2**	**RMSE**	**MAE**
Case Study 1: COVID-19 Vaccine Intention				
Virginia - 2021 (Direct Survey-Based Approach)	0.737	0.543	11.790	8.991
Virginia - 2021 (Regression Approach)	0.690	0.476	13.829	10.672
Virginia - 2021 (Probabilistic Approach)	0.657	0.431	12.997	9.966
Case Study 2: Depression				
Virginia - 2021	0.759	0.576	4.533	4.093
New Jersey - 2021	0.531	0.282	3.349	2.947
Tennessee - 2021	0.744	0.553	8.775	8.458
Case Study 3: Smoking				
Virginia - 2021	0.762	0.580	3.763	2.890
Virginia - 2022	0.816	0.665	4.103	3.206

3.1 Case Study 1: Comparing Approaches with COVID-19 Vaccine Intention

Overall accuracy. The moderate statistical measures of r and R^2 (Table 1) indicate that synthetic populations incorporating COVID-19 vaccine intention through the direct survey-based, regression, and probabilistic approaches reasonably reflect the observed vaccine uptake in Virginia as of December 2021. The direct survey-based approach yields the highest values (r = 0.74; R^2 = 0.54), while the regression (r = 0.69; R^2 = 0.48) and probabilistic approaches (r = 0.66; R^2 = 0.43) perform similarly. The direct survey-based approach achieves the lowest RMSE (11.79) and MAE (8.99), compared to the regression-based approach (RMSE = 13.83; MAE = 10.67) and the probabilistic approach (RMSE = 13.00; MAE = 9.97).

Preservation of statistical relationships. Logistic regression analysis of the UAS-UCA survey (Table 2) reveals that individuals aged 65+, male, with a bachelor's degree, or with an income above $100,000 are more likely to get the COVID-19 vaccine, with education and age being the strongest predictors. Conversely, Black individuals exhibit lower vaccine intention rates. The synthetic populations generated using each of the three approaches produce regression coefficients closely matching those from the UAS-UCA survey, effectively preserving real-world statistical relationships. Notably, the regression-based synthetic population has identical coefficients to the UAS-UCA survey, as vaccine intention was directly predicted from the survey. Despite this, the direct survey-based approach produces coefficients that align more closely with those from the UAS-UCA survey compared to the probabilistic approach.

Spatial heterogeneity. Figure 1 shows clusters and outliers of vaccine-related behavior identified by the Local Moran's I method across 1,592 Virginia census tracts with available data. A 'High-High Cluster' (light pink) indicates census tracts with high vaccine-related behavior surrounded by similarly high-rate tracts, while a 'Low-Low Cluster' (light blue) represents tracts with low vaccine behavior surrounded by others with low rates. Outlier tracts include 'High-Low Outliers' (bright red), where high-vaccine tracts are

Table 2: Coefficients comparing relationships between demographics and health outcomes from the input surveys and synthesized populations. All shown are significant with 99% confidence. Education was excluded from case study 2, and gender from case study 3, due to lack of significance in the survey data.

Case Study 1: COVID-19 Vaccine Intention				
Variable: Descriptor	***UAS-UCA***	***Direct Survey-Based***	***Regression***	***Probabilistic***
Age: 65 and over	0.558	0.560	0.558	0.480
Gender: Male	0.249	0.262	0.249	0.337
Education: Bachelor's or higher	0.763	0.885	0.763	0.790
Income: Greater than $100,000	0.350	0.265	0.350	0.205
Race/Ethnicity: Black	-0.280	-0.215	-0.280	-0.151
Intercept	0.136	0.081	0.136	0.143
Case Study 2: Depression				
Variable: Descriptor	***UAS-UCA***	***VA-2021***	***NJ-2021***	***TN-2021***
Age: 65 and over	-0.322	-0.295	-0.240	-0.415
Gender: Male	-0.837	-0.771	-0.833	-0.866
Income: Greater than $100,000	-0.624	-0.672	-0.589	-0.808
Race/Ethnicity: Black	-0.564	-0.646	-0.567	-0.839
Intercept	-0.722	-0.795	-0.860	-0.599
Case Study 3: Smoking				
Variable: Descriptor	***UAS-UCA***	***VA-2021***	***UAS-DUS***	***VA-2022***
Age: 65 and over	-1.048	-0.760	-1.017	-0.788
Education: Bachelor's or higher	-1.028	-1.030	-1.138	-1.264
Income: Greater than $100,000	-1.020	-0.772	-1.316	-0.987
Race/Ethnicity: White	0.576	0.361	0.575	0.397
Intercept	-1.710	-1.723	-1.680	-1.775

surrounded by low-vaccine areas, and 'Low-High Outliers' (bright blue), where the reverse occurs. Census tracts with no significant spatial relationship are shown in light yellow, and those lacking population or vaccine data appear in grey.

Figure 1A shows observed vaccine uptake as of December 2021, with lower uptake in western and southwestern Virginia and high-uptake clusters in Northern Virginia and central areas of the state. The remaining areas show mixed uptake, leading to outlier formation. Spatial patterns of COVID-19 vaccine intent from synthetic populations generated using the direct survey-based (Figure 1B), regression (Figure 1C), and probabilistic approaches (Figure 1D) generally align with observed vaccine uptake. Minor differences emerge, particularly in the clustering in central Virginia and the presence of outliers in southeastern Virginia.

3.2 Case Study 2: Demonstrating Spatial Adaptability with Depression

Overall accuracy. The evaluation metrics presented in Table 1 show better performance for synthesizing populations with depression compared to vaccine intention. Specifically, r and R^2 values are moderately high for the synthesized populations representing Virginia (r = 0.76; R^2 = 0.58) and Tennessee (r = 0.75; R^2 = 0.55), but the results are notably weaker for New Jersey (r = 0.53; R^2 = 0.28). Despite this, the New Jersey population results in the lowest RMSE and MAE values (RMSE = 3.35, MAE = 2.95) compared to the populations for Virginia (RMSE = 4.53, MAE = 4.09) and Tennessee (RMSE = 8.78, MAE = 8.46). This suggests that while the synthesized population for New Jersey does not capture the variance in depression from the CDC's PLACES estimates as well, the predictions are closer in magnitude. In contrast, the Tennessee population better captures the variance in depression but produces greater error magnitudes.

Preservation of statistical relationships. The logistic regression results from the UAS-UCA survey indicate that individuals who are aged 65+, male, have an income greater than $100,000, or Black are less likely to be diagnosed with depression. These demographics were relatively strong predictors, except for age, which consistently showed the weakest association. Education was not a significant predictor. Similar

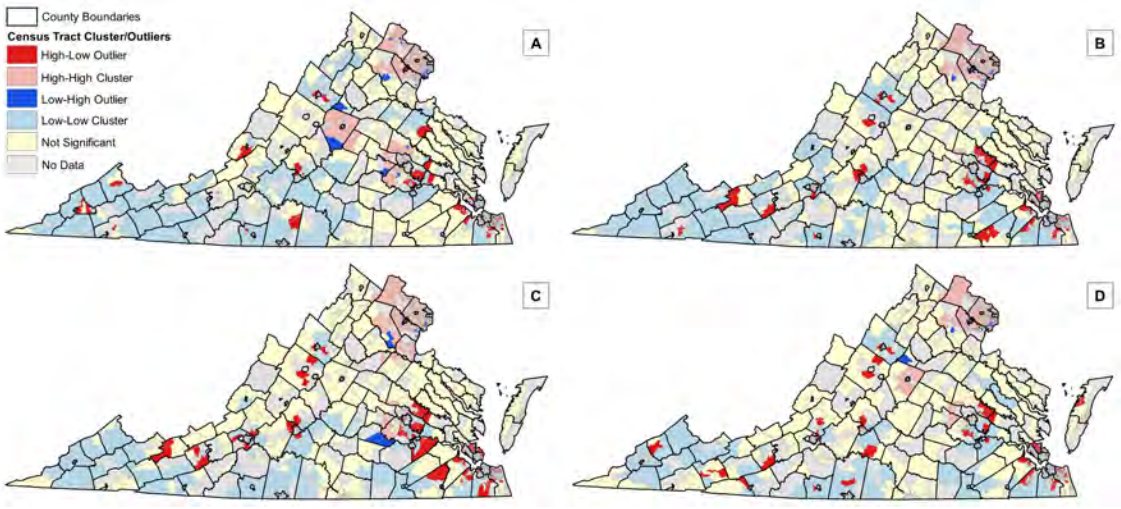

Figure 1: Clusters and outliers across Virginia census tracts of observed vaccine uptake as of Dec. 2021 (A) and aggregated synthesized population (%) with COVID-19 vaccine intention using the direct survey-based (B), regression (C), and probabilistic (D) approaches.

statistical relationships were observed in the regression analyses for the synthetic populations representing Virginia (R^2 = 0.04) , New Jersey (R^2 = 0.04) , and Tennessee (R^2 = 0.06) . In these analyses, gender consistently emerged as the strongest predictor, while age remained the weakest. Minor differences in coefficient values were noted across states.

Spatial heterogeneity. The clusters and outliers maps in Figure 2 illustrate spatial patterns of diagnosed depression in the synthesized populations compared to the CDC's PLACES estimates for Virginia, Tennessee, and New Jersey. In Virginia's synthetic population (Figure 2B), high depression rates cluster in the western half of the state, while Northern Virginia shows lower rates. These patterns generally align with the PLACES data (Figure 2A), with minor discrepancies in the southeast. For Tennessee, the synthetic population shows clustering of high depression rates in the eastern half of the state (Figure 2D), with less depression seen in the central region and southwestern counties. Similar patterns appear in the PLACES data (Figure 2C), though there is slightly less 'high-high' clustering in the middle and southern parts of the state. The synthetic population for New Jersey (Figure 2F) shows higher depression rates in the northwest, along the eastern coast, and in the south, while lower rates are concentrated in the northeast and central regions. This mirrors the PLACES data (Figure 2E), although there is slightly more clustering of high depression rates in the northwestern county and along the southwestern border.

3.3 Case Study 3: Modeling Temporal Changes with Smoking

Overall accuracy. Similar to the populations synthesized for depression, those generated for smoking behaviors in Virginia demonstrated moderate evaluation performance. Specifically, the 2021 smoking behavior population adequately explained real-world variance and exhibited relatively low magnitude of error, with r = 0.76, R^2 = 0.58, RMSE = 3.76, and MAE = 2.89. However, the population synthesized for 2022 showed an improved ability to capture the relationship and variance in real-world smoking behaviors (r = 0.82, R^2 = 0.67), but had slightly higher error values (RMSE = 4.10, MAE = 3.21).

Preservation of statistical relationships. Table 2 summarizes the coefficients explaining smoking behavior across the surveys and populations generated for 2021 and 2022. Gender was not included since it was not a statistically significant predictor of smoking. In all cases, the coefficients reveal consistent relationships. Specifically, individuals who are 65 or older, hold a bachelor's degree, or have an income greater than $100,000 are associated with lower likelihoods of smoking, as indicated by the negative

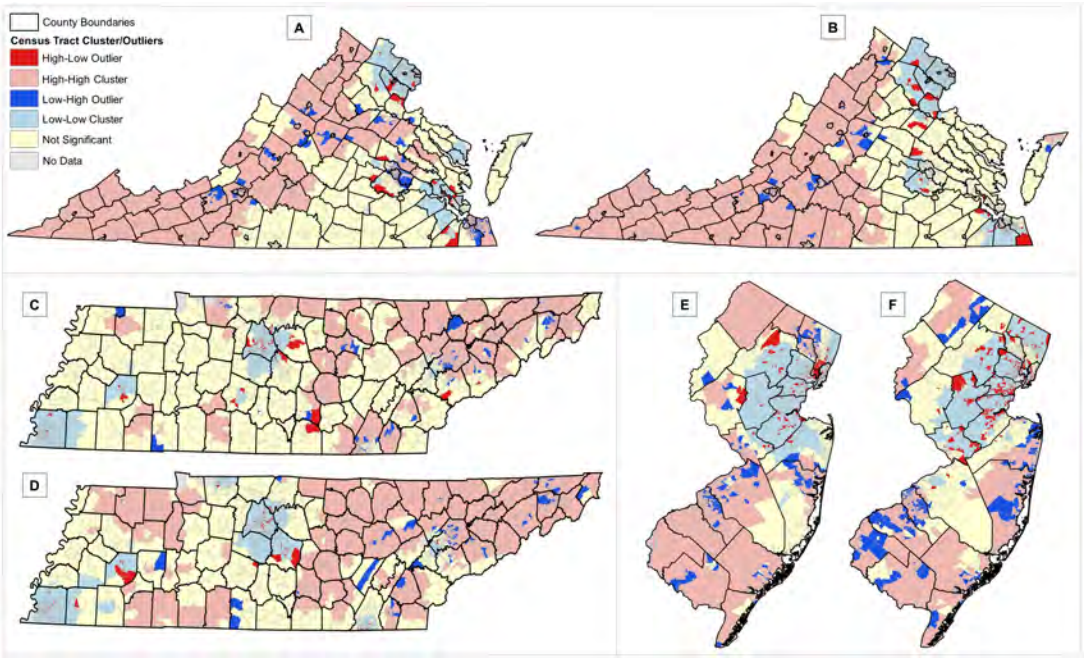

Figure 2: Clusters and outliers in the aggregated population (%) with diagnosed depression across census tracts: (A) VA CDC PLACES estimates, (B) VA synthesized population, (C) TN CDC PLACES estimates, (D) TN synthesized population , (E) NJ CDC PLACES estimates, and (F) NJ synthesized population.

coefficients. The relationships between race/ethnicity and smoking shows slight variation, but positive coefficients consistently suggest a higher likelihood of smoking among White individuals in each dataset, with synthetic populations showing lower magnitudes. A similar pattern is observed across each survey and its corresponding synthetic population, where the survey datasets generally exhibit larger magnitudes for age and income coefficients, with income showing the largest difference, suggesting a stronger inverse relationship between high income and smoking in these models. Additionally, the 2022 survey dataset and the synthetic population for 2022 show stronger negative associations for education and income compared to the 2021 datasets. Interestingly, each regression model exhibited greater variation in R^2 values compared to the previous case studies (UAS-UCA = 0.08, VA-2021 = 0.05, UAS-DUS = 0.10, VA-2022 = 0.06).

Spatial heterogeneity. The CDC's PLACES estimates indicate that high-smoking clusters are concentrated in the south region of Virginia, while lower-smoking prevalence is more common in northern Virginia. Similar spatial patterns emerge in both the 2021 and 2022 synthetic populations. A spatial comparison between the two reveals subtle shifts in these patterns, suggesting changes in smoking behavior over time. These temporal changes are visualized in Figure 3, which highlights shifts in smoking behavior from 2021 to 2022 through clusters and outliers. The 'high-high' clustering near the south-central border and the Eastern Shore of Virginia indicates areas with a significant increase in smoking, while the 'low-low' clustering in Northern Virginia reflects a reduction in smoking. Additionally, scattered 'high-low' outliers, such as those in the central-west region of the state, show census tracts where smoking increased despite being surrounded by areas with the opposite trend.

4 DISCUSSION AND CONCLUSION

This study's results provide evidence that the direct survey-based approach introduced in Von Hoene et al. (2025) can be used to generate synthetic populations across different health outcomes, locations, and points in time. The approach offers a combination of overall accuracy, preservation of statistical relationships between health outcomes and demographic variables, and close reflection of the true spatial heterogeneity.

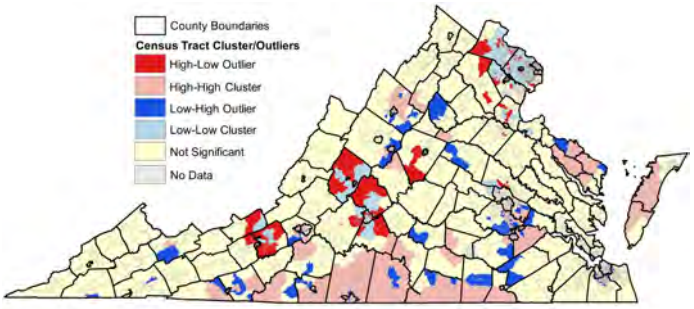

Figure 3: Clusters and outliers of change (%) in smoking in synthesized populations, 2021 to 2022.

Beyond accuracy, the direct survey-based approach offers significant advantages in computational efficiency and ease of use, as it directly integrates survey data into the population synthesis process without requiring regression models or iterative probability calculations. This supports researchers already conducting synthetic population generation and seeking a more streamlined method for their own work. It also addresses the growing need for synthetic populations in social simulations by enabling the creation of tailored populations for domain-specific modeling across a wide range of research applications.

Quantitative evaluation (Table 1) suggests that our direct survey-based approach may outperform traditional regression and probabilistic methods. This improved performance likely stems from the direct integration of surveys into the population synthesis process. In contrast, the regression and probabilistic methods must learn from data to assign outcomes to a base population, leaving more room for generalization and error. Although, all three methods assume a relationship between demographics and health outcomes. In practice, these relationships may vary across space or be influenced by additional unobserved factors.

Notably, we find that the direct survey-based and probabilistic approaches converge towards a similar synthetic population, particularly as the number of demographic combinations (hereafter referred to as strata) considered in the probabilistic approach increases. This is shown in Figure 4, where the blue line reflects increasing similarity between the two approaches (measured with the R^2 metric), and the red line shows the accuracy of the probabilistic approach relative to the validation data. At first, increasing the number of strata considered in the probabilistic approach —first by incorporating education, then age, and subsequently income—leads to results that gradually converge with those of the population synthesized using the direct survey-based approach (from an R^2 of 0.53 to 0.76). This also improves the accuracy of the synthetic population relative to the validation data (from an R^2 of 0.42 to 0.51). However, given the smaller influence of race and gender on vaccine intention, using strata that consider race and gender slow this convergence (from an R^2 of 0.76 to 0.77) and decrease the accuracy of the populations when compared to real data (from an R^2 of 0.51 to 0.43). This highlights the importance of carefully selecting the strata for which to generate probabilities in population synthesis approaches.

In general, a key limitation of population synthesis is that it depends on the quality of the input data used. In prior studies, we observed that over-representation of an outcome in the national survey data can inflate this outcome in the synthetic population (Von Hoene et al. 2025). In this study, we found that our approach was more successful when the study area was more aligned with the nationally representative data. For example, when both the survey and Virginia spatial data report that 20% of the population has diagnosed depression. However, more research is needed to understand and mitigate the effect of bias on the synthetic populations. Additionally, the validation results depend on the quality of the available data. For instance, our findings on depression and smoking behaviors are likely more accurate than those on COVID-19 vaccine intent due to greater data availability, whereas COVID-19 vaccine data is often limited or prone to misreporting.

In addition to enhancing population synthesis for social simulations like agent-based and microsimulations, the direct survey-based approach supports the creation of public health datasets at finer spatial (e.g., census tract) and temporal (e.g., monthly or yearly) resolutions. This addresses the limitations of

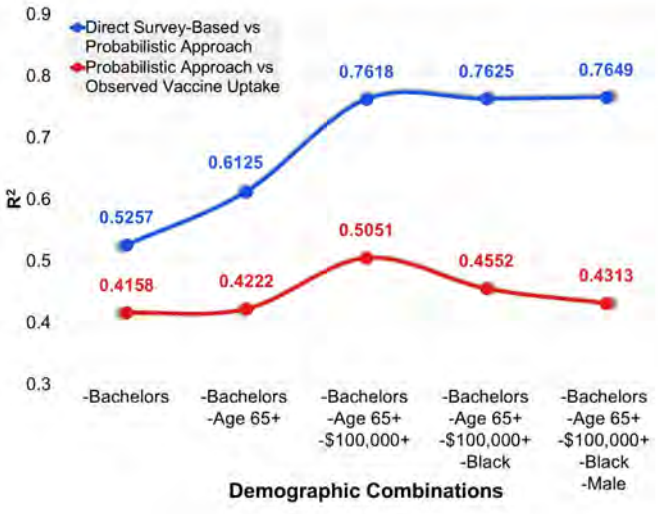

Figure 4: Computed R^2 values for each iteration of the probabilistic approach, comparing (1) simulated COVID-19 vaccine intent from the direct survey-based and probabilistic approach (blue) and (2) COVID-19 vaccine intent from the probabilistic approach with actual vaccine uptake in Virginia (Dec. 2021, red).

traditional data sources, such as five-year surveys, which are costly to produce and may lack the granularity needed for localized health research. While we generate synthetic populations with specific outcomes like vaccine intention, depression, and smoking so that we can validate, this method can also be used to generate populations with attitudes and perceptions and other relevant attributes (see Von Hoene et al., 2025). These small area estimates produced by our approach offer a resource for examining demographic and spatial disparities in health, ultimately supporting the design of geographically targeted public health strategies.

In summary, we demonstrate how direct survey-based population synthesis can be used for various public health outcomes at different locations and points in time. To our knowledge, this method has been the first to integrate domain-specific surveys directly into the population synthesis process to generate populations with relevant public health attributes. Future work could expand this approach by combining multiple surveys to create populations with a more comprehensive set of attributes tailored to specific study applications. For instance, while the UAS-UCA includes significant behavioral measures, it does not capture factors relating to financial or geographic barriers. By combining datasets, we can address such gaps and create more holistic populations for social simulations, ultimately enhancing decision-making. We also encourage other researchers to apply this approach to study applications beyond public health, with the aim of improving the accuracy of simulating human behavior in large-scale social models.

ACKNOWLEDGMENTS

This research was funded by the National Science Foundation (Awards #2109647 and #2302970). The content of this paper is the responsibility of the authors and does not necessarily represent the official views of UAS. The collection of UAS data related to COVID-19 is supported in part by the Bill & Melinda Gates Foundation, grant U01AG054580 from the National Institute on Aging, and additional funding sources.

REFERENCES

Akhtar-Danesh, N., and J. Landeen. 2007. "Relation Between Depression and Sociodemographic Factors". *International Journal of Mental Health Systems* 1(1):4 https://doi.org/10.1186/1752-4458-1-4.

AlShurman, B. A., A. F. Khan, C. Mac, M. Majeed, and Z. A. Butt. 2021. "What Demographic, Social, and Contextual Factors Influence the Intention to Use COVID-19 Vaccines: A Scoping Review". *International Journal of Environmental Research and Public Health* 18(17):9342 https://doi.org/10.3390/ijerph18179342.

Anselin, L. 1995. "Local Indicators of Spatial Association—LISA". *Geographical Analysis* 27(2):93–115.

Axelrod, R. 2007. "Simulation in Social Sciences". In *Handbook of Research on Nature-Inspired Computing for Economics and Management*, 90–100. IGI Global.

Bigi, F., T. H. Rashidi, and F. Viti. 2024. "Synthetic Population: A Reliable Framework for Analysis for Agent-Based Modeling in Mobility". *Transportation Research Record* 2678(11):1–15 https://doi.org/10.1177/0361198124123965.

Chapuis, K., P. Taillandier, and A. Drogoul. 2022. "Generation of Synthetic Populations in Social Simulations: A Review of Methods and Practices". *Journal of Artificial Societies and Social Simulation* 25(2):6 https://doi.org/10.18564/jasss.4752.

Centers for Disease Control and Prevention 2024. "Behavioral Risk Factor Surveillance System Survey".

Centers for Disease Control and Prevention 2025. "CDC PLACES: Local Data for Better Health".

Garrett, B. E., B. N. Martell, R. S. Caraballo, and B. A. King. 2019. "Socioeconomic Differences in Cigarette Smoking Among Sociodemographic Groups". *Prev Chronic Dis* 16:E74 https://doi.org/10.5888/pcd16.180553.

Huang, Z., and P. Williamson. 2001. "A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small-Area Microdata". Technical report, Department of Geography, University of Liverpool.

Knight, J., S. Wells, R. Marshall, D. Exeter, and R. Jackson. 2017. "Developing a Synthetic National Population to Investigate the Impact of Different Cardiovascular Disease Risk Management Strategies: A Derivation and Validation Study". *PLOS One* 12(4):e0173170 https://doi.org/10.1371/journal.pone.0173170.

Lovelace, R., and D. Ballas. 2013. "'Truncate, Replicate, Sample': A Method for Creating Integer Weights for Spatial Microsimulation". *Computers, Environment and Urban Systems* 41 https://doi.org/10.1016/j.compenvurbsys.2013.03.004.

Nicolaie, M. A., K. Füssenich, C. Ameling, and H. C. Boshuizen. 2023. "Constructing Synthetic Populations in the Age of Big Data". *Population Health Metrics* 21(1):19 https://doi.org/10.1186/s12963-023-00319-5.

Pandey, A., M. C. Fitzpatrick, S. M. Moghadas, T. N. Vilches, C. Ko, A. Vasan *et al*. 2023. "Modelling the Impact of a High-Uptake Bivalent Booster Scenario on the COVID-19 Burden and Healthcare Costs in New York City". *The Lancet Regional Health–Americas* 24 https://doi.org/10.1016/j.lana.2023.100555.

Prem, K., A. R. Cook, and M. Jit. 2017. "Projecting Social Contact Matrices in 152 Countries Using Contact Surveys and Demographic Data". *PLOS Computational Biology* 13(9):e1005697 https://doi.org/10.1371/journal.pcbi.1005697.

University of Southern California 2025. "Understanding America Study".

U.S. Census Bureau 2025. "American Community Survey (ACS)".

Von Hoene, E., A. Roess, H. Kavak, and T. Anderson. 2025, 03. "Synthetic Population Generation with Public Health Characteristics for Spatial Agent-Based Models". *PLOS Computational Biology* 21(3):1–22 https://doi.org/10.1371/journal.pcbi.1012439.

Zhang, X., J. Holt, S. Yun, H. Lu, K. Greenlund, and J. Croft. 2015. "Validation of Multilevel Regression and Poststratification Methodology for Small Area Estimation of Health Indicators from the Behavioral Risk Factor Surveillance System". *American Journal of Epidemiology* 182(2):127–137 https://doi.org/10.1093/aje/kwv002.

Zhu, K., L. Yin, K. Liu, J. Liu, Y. Shi, X. Li, *et al*. 2024. "Generating Synthetic Population for Simulating the Spatiotemporal Dynamics of Epidemics". *PLOS Computational Biology* 20(2):e1011810 https://doi.org/10.1371/journal.pcbi.1011810.

AUTHOR BIOGRAPHIES

EMMA VON HOENE is a Ph.D. student in the Department of Geography and Geoinformation Science at George Mason University in Fairfax, Virginia, USA. She holds a Master of Science in Geoinformatics and Geospatial Intelligence from GMU. Her research focuses on data-driven modeling and geospatial analysis. Her email address is evonhoen@gmu.edu.

AANYA GUPTA is a high school student at Thomas Jefferson High School for Science and Technology in Alexandria, Virginia, USA. Her research interests include applications of computer science and computational modeling for public health. Her e-mail address is 2026agupta@tjhsst.edu.

HAMDI KAVAK is currently Associate Professor in the Department of Computational and Data Sciences and Co-Director of Center for Social Complexity at George Mason University. His research combines data science with modeling and simulation to investigate challenges in urban and social systems. His email address is hkavak@gmu.edu and his website is https://hamdikavak.com/.

AMIRA ROESS is a Professor of Global Health and Epidemiology at George Mason University. Her research focuses on emergence and transmission of infectious diseases and the development of multi-disciplinary and multi-species field research methods. Her email address is aroess@gmu.edu and her website is https://publichealth.gmu.edu/profiles/aroess.

TAYLOR ANDERSON is an Associate Professor in the Department of Geography and Geoinformation Science at George Mason University. Her research focuses on modeling the spread of diseases in human and ecological systems. Her e-mail address is tander6@gmu.edu and her website is https://science.gmu.edu/directory/taylor-anderson.