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] , Taylor Anderson[1]

1. Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA 22030

2. Thomas Jefferson High School for Science and Technology, Alexandria, VA 22312

3. Department of Computational and Data Sciences, George Mason University, Fairfax, VA 22030

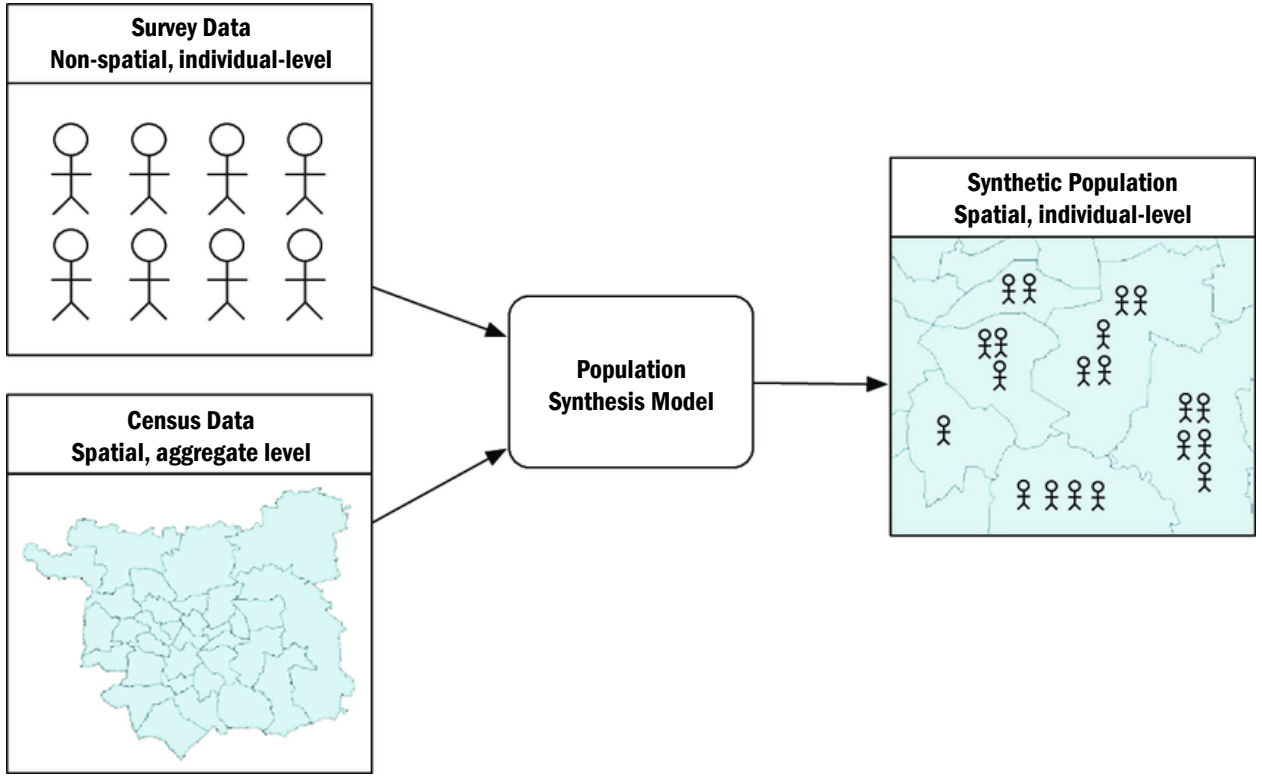
4. Department of Global and Community Health, George Mason University, Fairfax, VA 22030

Data Analysis in Healthcare Applications
2025 Winter Simulation Conference
December 9th, 2025

GEORGE MASON UNIVERSITY

Introduction and Background

- **Synthetic population generation:** the process of creating representative artificial datasets of human individuals or households
  - Critical input for social simulations (e.g., ABMs), where outcomes are dependent on the heterogeneous characteristics of the population
  - Many methods, such as **Iterative Proportional Fitting (IPF),** to synthesize populations with socio-demographic attributes
    - See *Chapuis et al. (2022)* for an overview



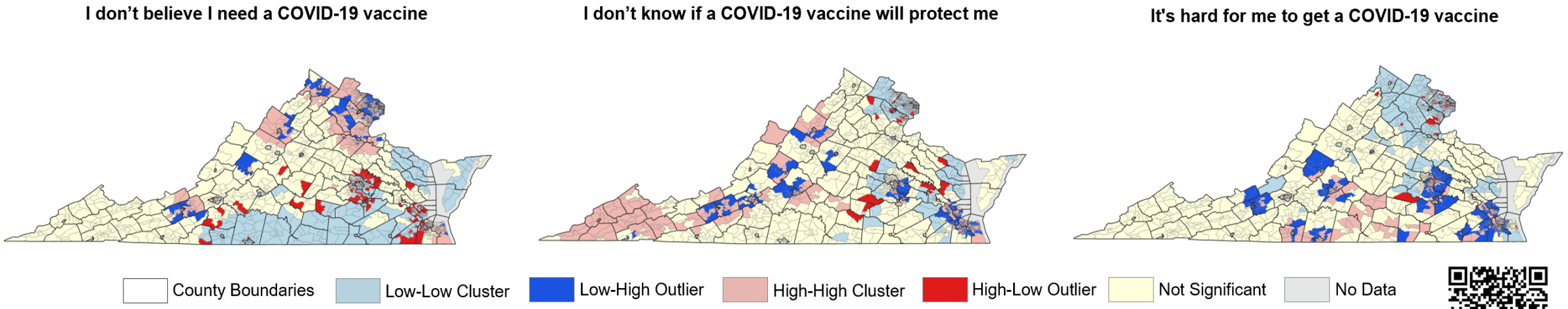
Schematic overview of synthetic population generation. Figure adapted from: Malleson et al. (2018).

*Chapuis et al. (2022). "**Generation of synthetic populations in social simulations: a review of methods and practices**." Journal of Artificial Societies and Social Simulation, 25(2).*

*Malleson et al. (2018). "**Place-based simulation modeling: agent-based modeling and virtual environments**."*

Introduction and Background

- Public health simulations require synthetic populations with health behaviors, or associated attitudes
- Traditionally, studies that synthesize populations with health-related attributes often use probabilistic- or regression-based approaches
  - Pandey et al. (2023) assign COVID-19 vaccine coverage based on age-specific **probabilities** of historic data
  - Nicolaie et al. (2023) use a sequence of **regression** models to add lifestyle and disease variables
- Von Hoene et al. (2025) propose a direct survey-based approach to initialize agents with health attitudes and behaviors



Clusters and outliers of synthetic populations generated with health attitudes, beliefs, and perceptions from Von Hoene et al. (2025).

QR code to Von Hoene et al. (2025).

*Pandey 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.*

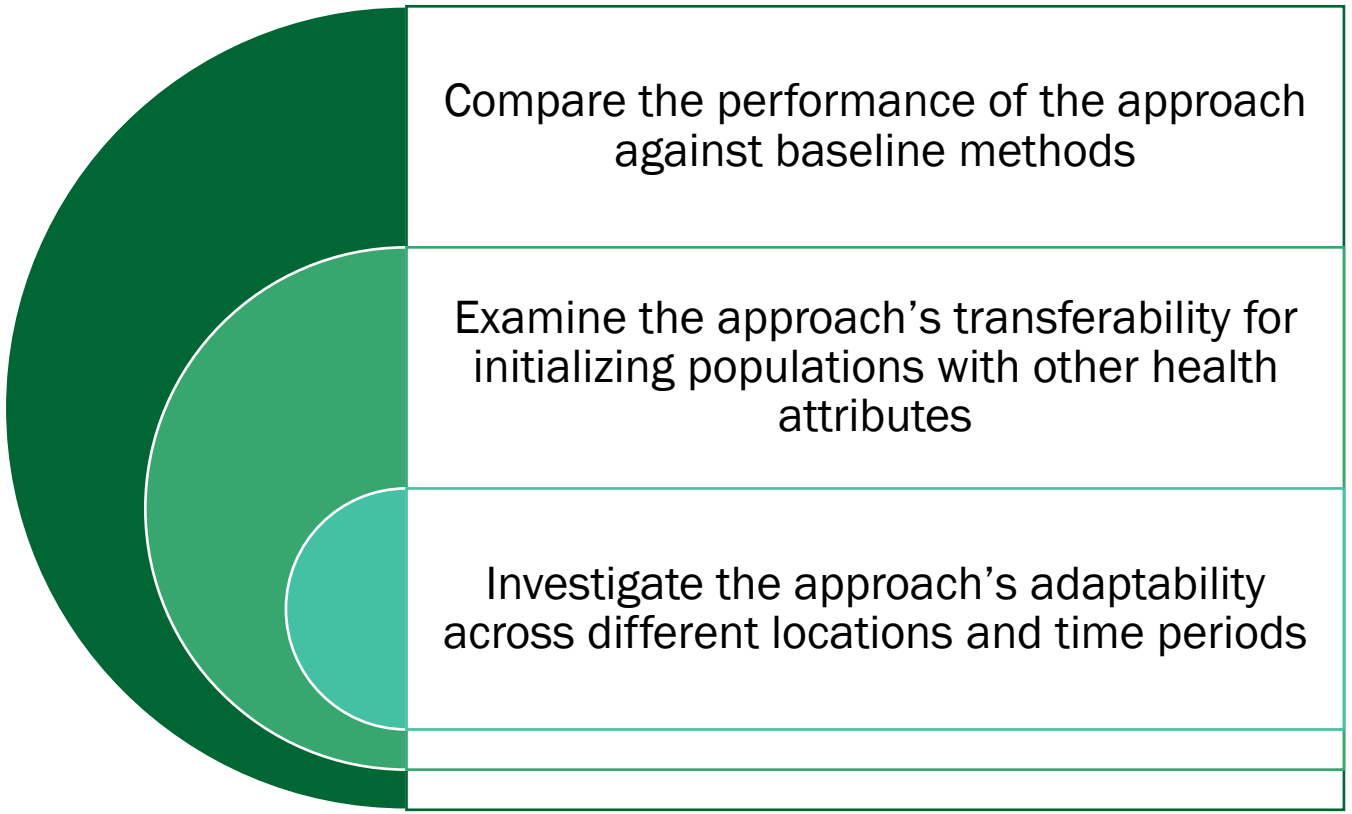
*Nicolaie et al. 2023. "**Constructing synthetic populations in the age of big data**". Population Health Metrics 21(1):19.*

*Von Hoene et al. (2025). **"Synthetic population generation with public health characteristics for spatial agent-based models."** PLOS Computational Biology, 21(3), e1012439.*

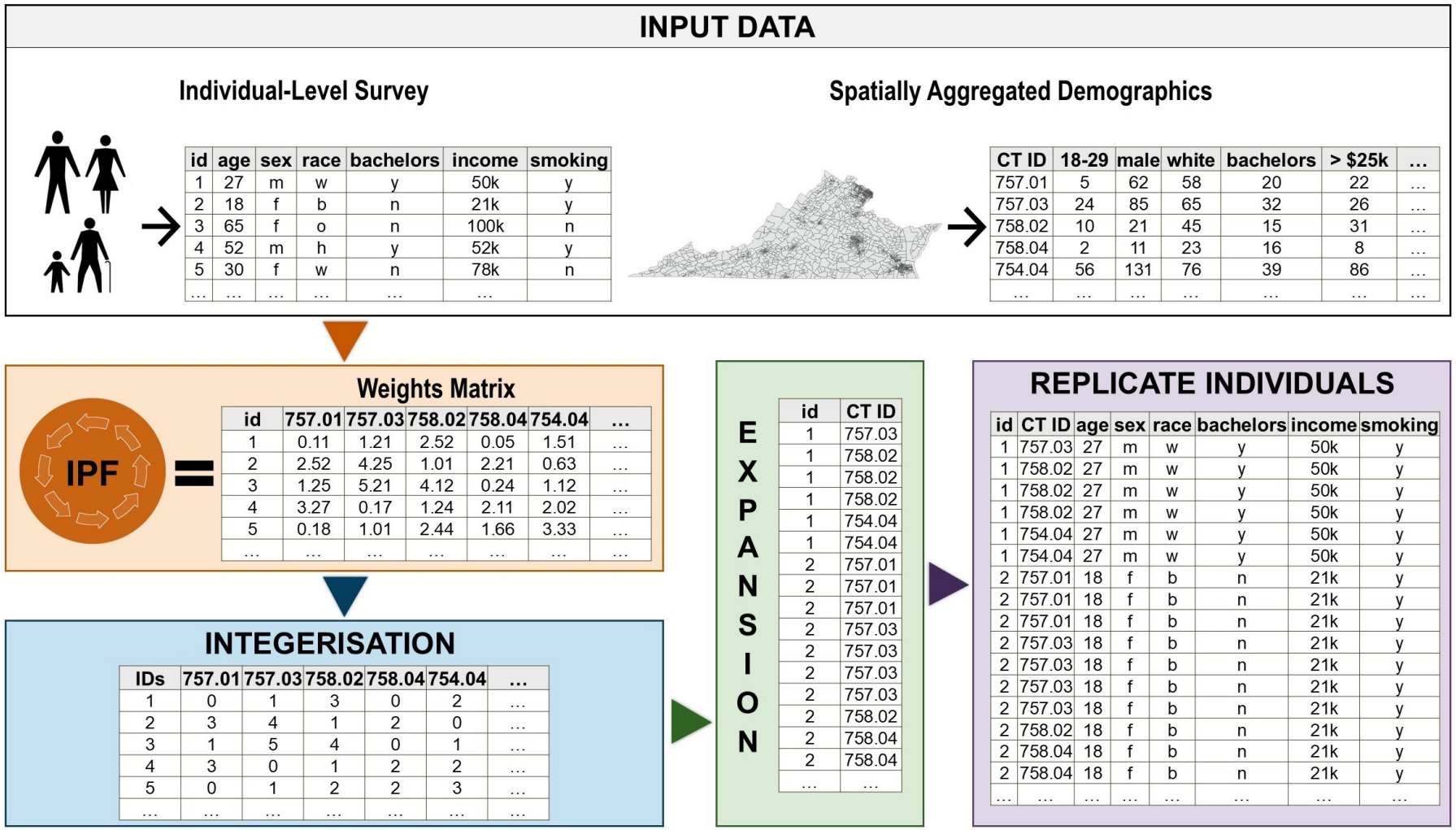
Objectives

Research Gap:

- The transferability of the **direct survey-based approach** remains unexplored.
- We address this by developing a series of synthetic populations that:



Methods: Population Synthesis



Overview of the direct survey-based synthetic population generation approach, from Von Hoene et al (2025).

Von Hoene et al. (2025). ***"Synthetic population generation with public health characteristics for spatial agent-based models."*** *PLOS Computational Biology, 21(3), e1012439.*

Methods: Data and Case Studies

Outline of each case study, its purpose, and study settings, along with corresponding input and validation datasets used.

Case Study/ Synthesized Outcome	Experiment Purpose	Time Period (s) and Study Area (s)	Disaggregated Individual-Level Survey Data	Spatially Aggregate Data	Validation Data
1. COVID-19 Vaccine Intention	Comparing Approaches	*2021:* Virginia	*March 2021*: Understanding America Survey – Understanding Coronavirus in America (**UAS-UCA**)	*2017-2021 5YR* **Census Tract** Estimates: US Census American Community Survey for age, race, gender, income, and education	*December 2021*: Virginia Department of Public Health Census Tract Level ***Observed*** Vaccine Uptake (Requested)
2. Depression	Demonstrating Spatial Adaptability	*2021:* Virginia, Tennessee and New Jersey			*2024 Release*: CDC PLACES Depression ***Model-Based*** Estimates
3. Smoking	Modeling Temporal Change	*2021 and 2022:* Virginia	**UAS-UCA** + *March 2022*: Understanding America Survey – Drug Use Supplement (**DUS**)		*2024 Release*: CDC PLACES Smoking ***Model-Based*** Estimates

Methods: Evaluation

We evaluated results for each synthesized population in terms of:

1. Overall accuracy

- Pearson's r
- R2
- RMSE
- MAE

2. Preservation of statistical relationships

- Logistic regression coefficients

3. Spatial heterogeneity

- Local Moran's I statistic

Results: Case Study 1

COVID-19 Vaccine Intention

Overall Accuracy:

Metrics comparing synthesized **COVID-19 vaccine intention** in **Virginia (2021)** from different approaches against observed vaccine uptake.

Population Synthesis Approach	Pearson r	R^2	RMSE	MAE
Direct Survey-Based Approach	0.737	0.543	11.790	8.991
Regression Approach	0.690	0.476	13.829	10.672
Probabilistic Approach	0.657	0.431	12.997	9.966

Preservation of Statistical Relationships:

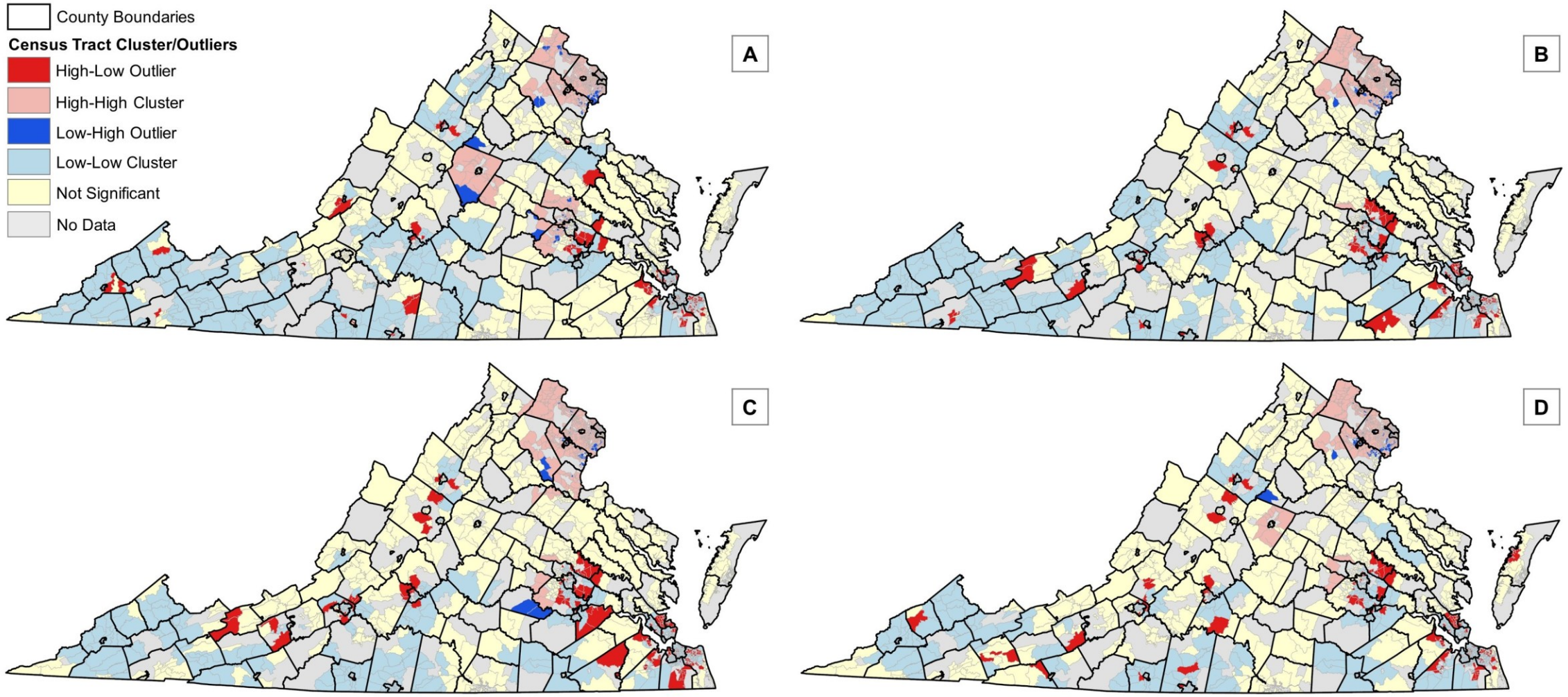
Logistic regression coefficients measuring the association between demographics and **COVID-19 vaccine intention** from the original UAS-UCA survey and synthesized **Virginia (2021) populations** across different approaches. *All variables are significant at 99% confidence.*

Variable: Descriptor	UAS-UCA	Direct Survey-Based	Regression*	Probabilistic
Age: 65+	0.558	0.560	0.558	0.480
Gender: Male	0.249	0.262	0.249	0.337
Education: Bachelor's degree or higher	0.763	0.885	0.763	0.790
Income: $100,000+	0.350	0.265	0.350	0.205
Race: Black	-0.280	-0.215	-0.280	-0.151
Intercept	0.136	0.081	0.136	0.143

** Coefficients are identical to UAS-UCA since this method uses logistic regression to synthesize COVID-19 vaccine intent*

Results: Case Study 1

COVID-19 Vaccine Intention



Spatial Heterogeneity:

Clusters and outliers of COVID-19 vaccine intention rates across Virginia census tracts as of December 2021 for A) observed vaccine uptake, and synthesized populations using the following approaches: B) Direct Survey-Based, C) Regression, D) Probabilistic.

Results: Case Study 2

Depression

Overall Accuracy:

Metrics comparing synthesized **depression (2021)** rates for different states against CDC PLACES estimates.

Location	Pearson r	R^2	RMSE	MAE
Virginia	0.759	0.576	4.533	4.093
New Jersey	0.531	0.282	3.349	2.947
Tennessee	0.744	0.553	8.775	8.458

Preservation of Statistical Relationships:

Logistic regression coefficients measuring the association between demographics and **depression (2021)** from the original UAS-UCA survey and synthesized populations for different states. Education is not included due to insignificance.

All variables are significant at 99% confidence.

Variable: Descriptor	UAS-UCA	Virginia	New Jersey	Tennessee
Age: 65+	-0.322	-0.295	-0.240	-0.415
Gender: Male	-0.837	-0.771	-0.833	-0.866
Income: $100,000+	-0.624	-0.672	-0.589	-0.808
Race: Black	-0.564	-0.646	-0.567	-0.839
Intercept	-0.722	-0.795	-0.860	-0.599

Results: Case Study 2

Depression

Spatial Heterogeneity: Clusters and outliers of **depression** across census tracts for:

A) Virginia CDC PLACES

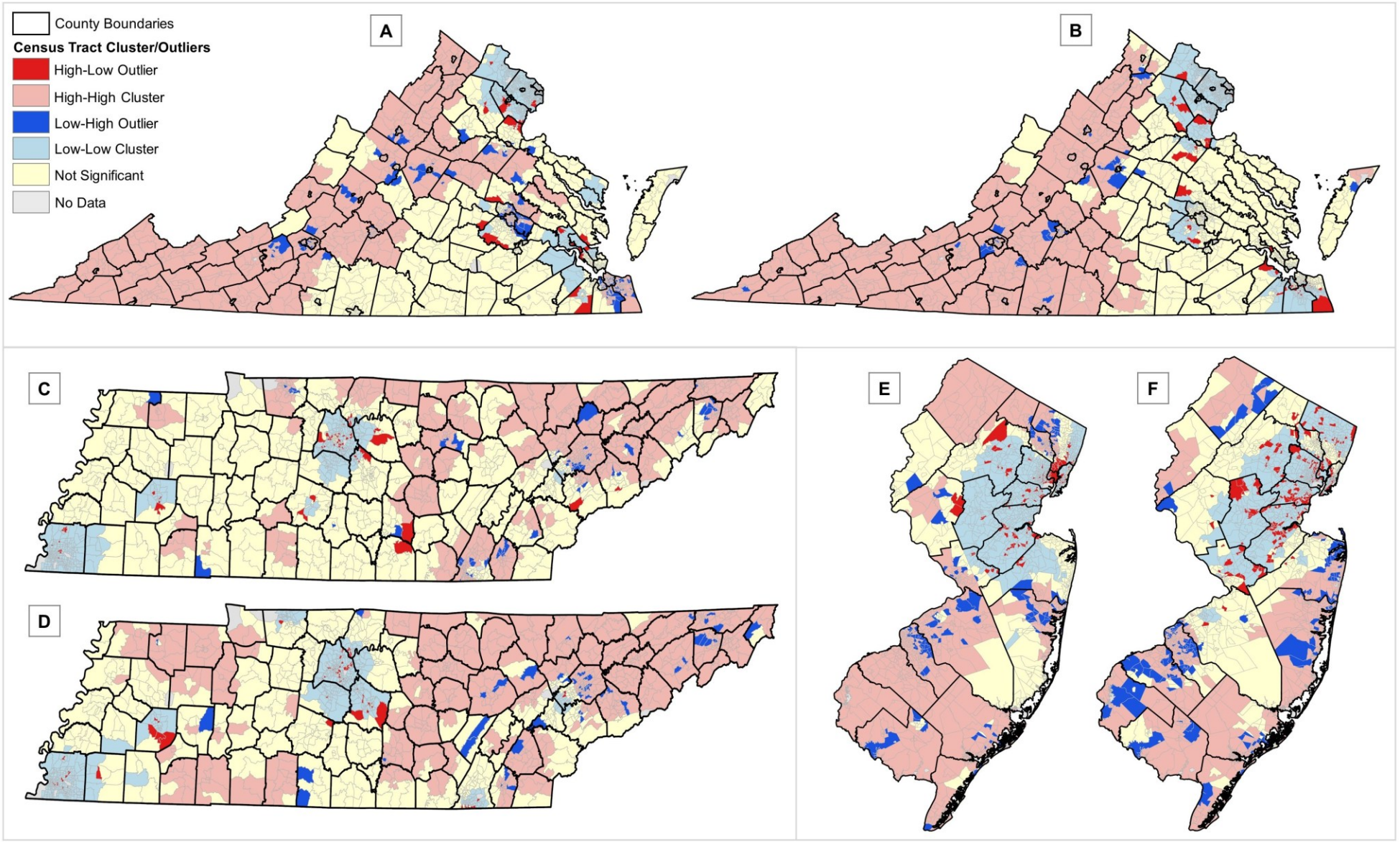
B) Virginia synthesized population

C) Tennessee CDC PLACES

D) Tennessee synthesized population

E) New Jersey CDC PLACES

F) New Jersey synthesized population



Results: Case Study 3

Smoking

Overall Accuracy:

Metrics comparing synthesized **Virginia** smoking rates from different time periods against CDC PLACES estimates.

Year	Pearson r	R^2	RMSE	MAE
2021	0.762	0.580	3.763	2.890
2022	0.816	0.665	4.103	3.206

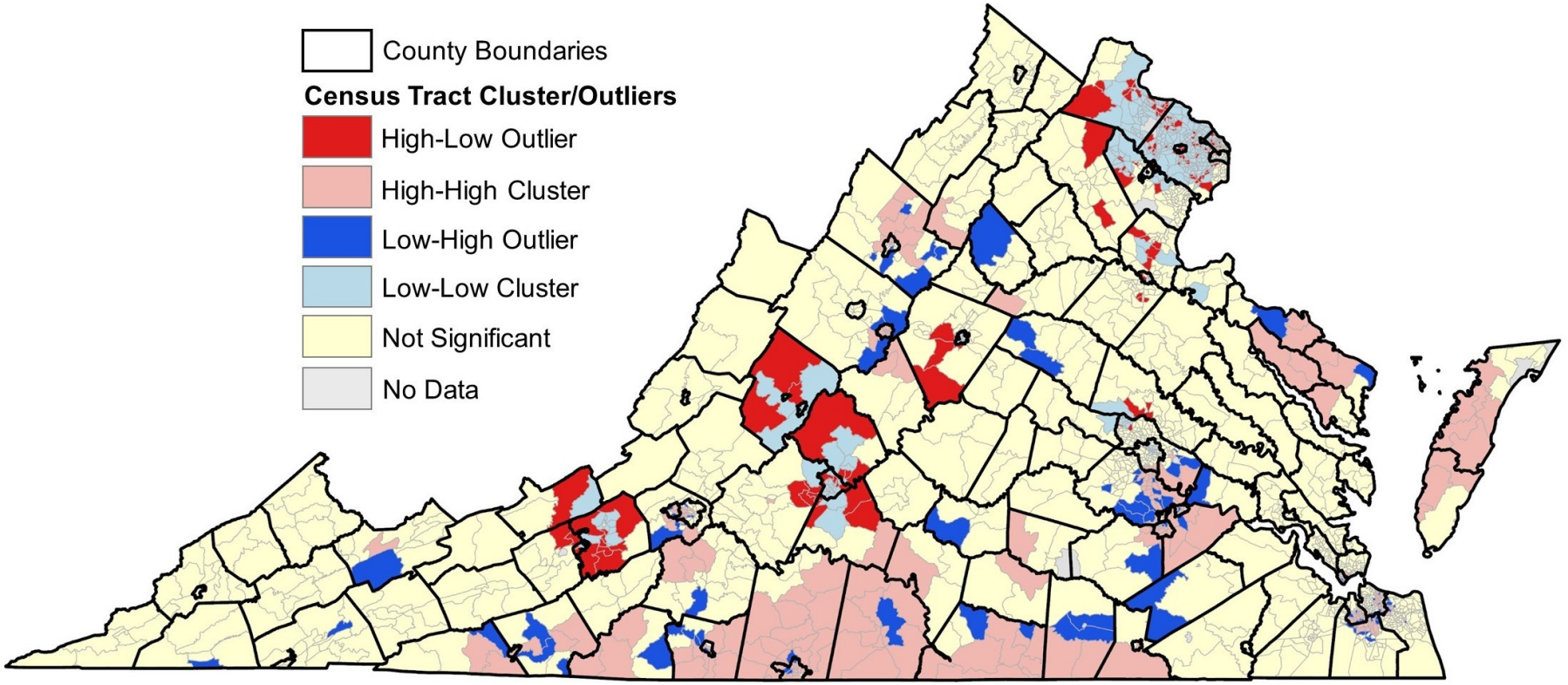
Preservation of Statistical Relationships:

Logistic regression coefficients measuring the association between demographics and **smoking** from the originally used UAS-UCA and UAS-UAS-DUS surveys and the synthesized Virginia populations for different time periods. Gender is not included due to insignificance.
All variables are significant at 99% confidence.

Variable: Descriptor	UAS-UCA	2021	UAS-DUS	2022
Age: 65+	-1.048	-0.760	-1.017	-0.788
Education: Bachelor's or higher	-1.028	-1.030	-1.138	-1.264
Income: $100,000+	-1.020	-0.772	-1.316	-0.987
Race: White	0.576	0.361	0.575	0.397
Intercept	-1.710	-1.723	-1.680	-1.775

Results: Case Study 3

Smoking



Spatial Heterogeneity:

Clusters and outliers of change in smoking rates between 2021 and 2022 in synthesized Virginia populations.

Discussion

The Direct Survey-Based Approach presented in Von Hoene et al. (2025):

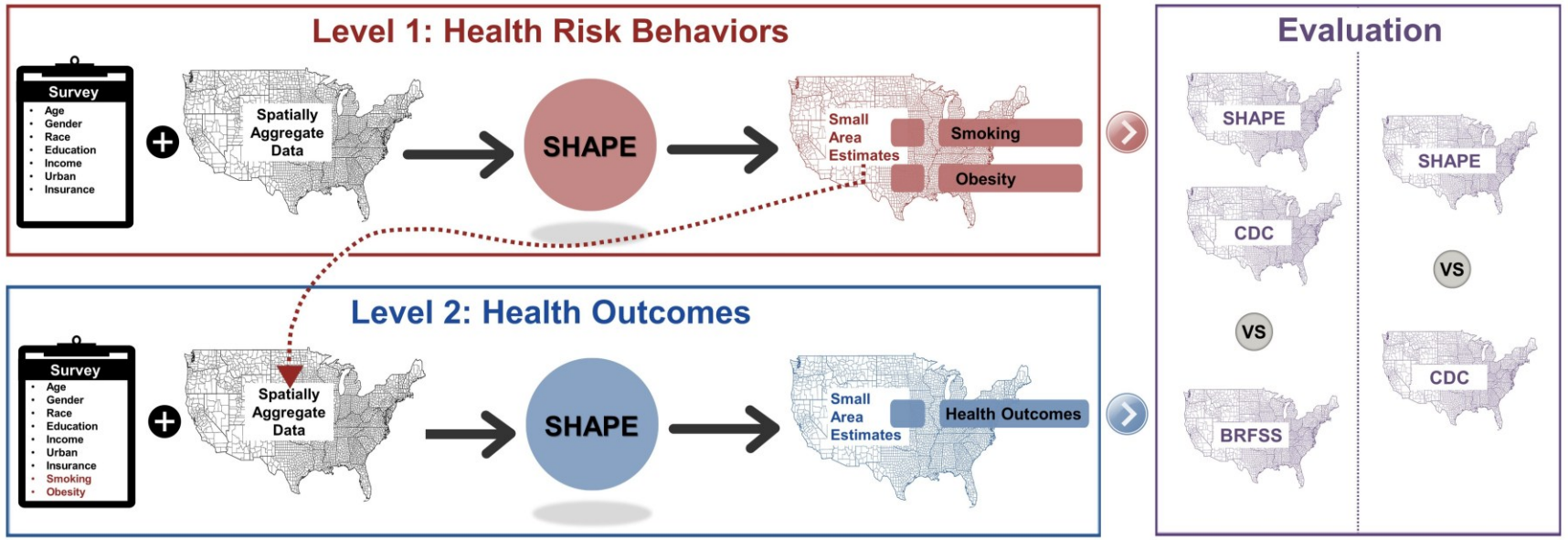
- Produces reliable synthetic populations across multiple health outcomes, locations and timeframes
- To our knowledge, it is the first method to directly integrate domain-specific surveys into the population synthesis process
- Limited based on quality of input and validation datasets → the case for all population synthesis approaches
- Offers advantages in computational efficiency and ease of use
- Outperforms other commonly used population synthesis approaches for health simulations
- Enhances population synthesis for social simulations and ***supports the development of small area estimates (SAE) for public health***

Von Hoene et al. (2025). ***"Synthetic population generation with public health characteristics for spatial agent-based models."*** *PLOS Computational Biology, 21(3), e1012439.*

Current Work: SHAPE

- **SHAPE (Spatial Health and Population Estimator):** an open, reproducible and flexible spatial microsimulation framework that leverages hierarchical IPF to generate public health small area estimates (SAEs)

- Applied SHAPE to generate SAEs for:
  - 2 health risk behaviors, 11 health outcomes
  - Florida (2019) and New York (2021)
  - Counties and census tracts
- Evaluated against BRFSS direct estimates & CDC PLACES

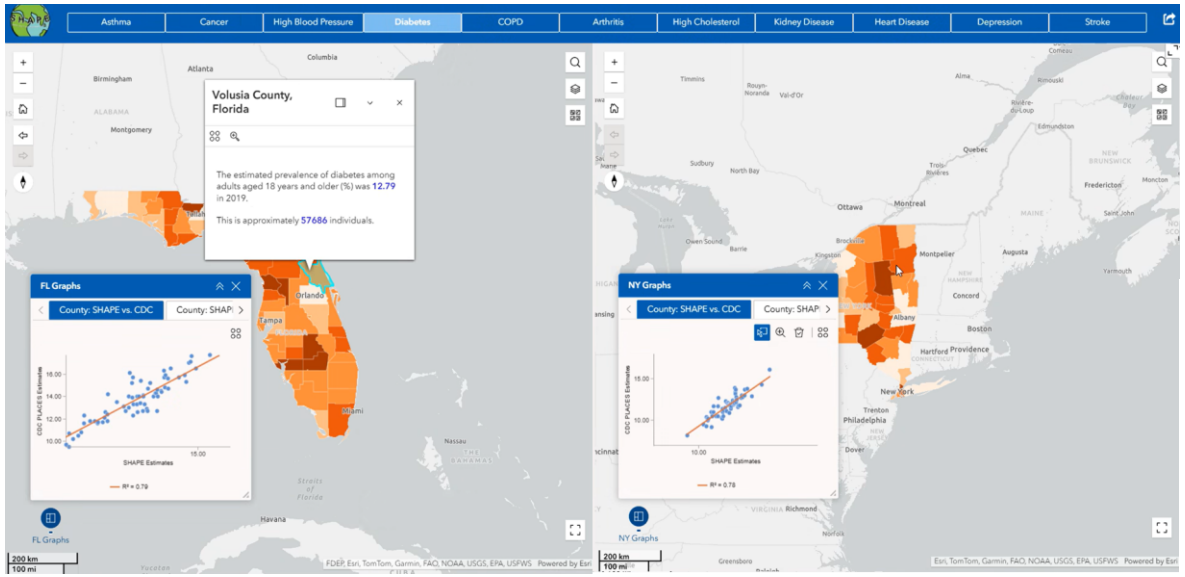


Overview of the two-leveled hierarchical IPF-based process used within the SHAPE framework.

Read more about SHAPE here:
tinyurl.com/fy8rk6u6

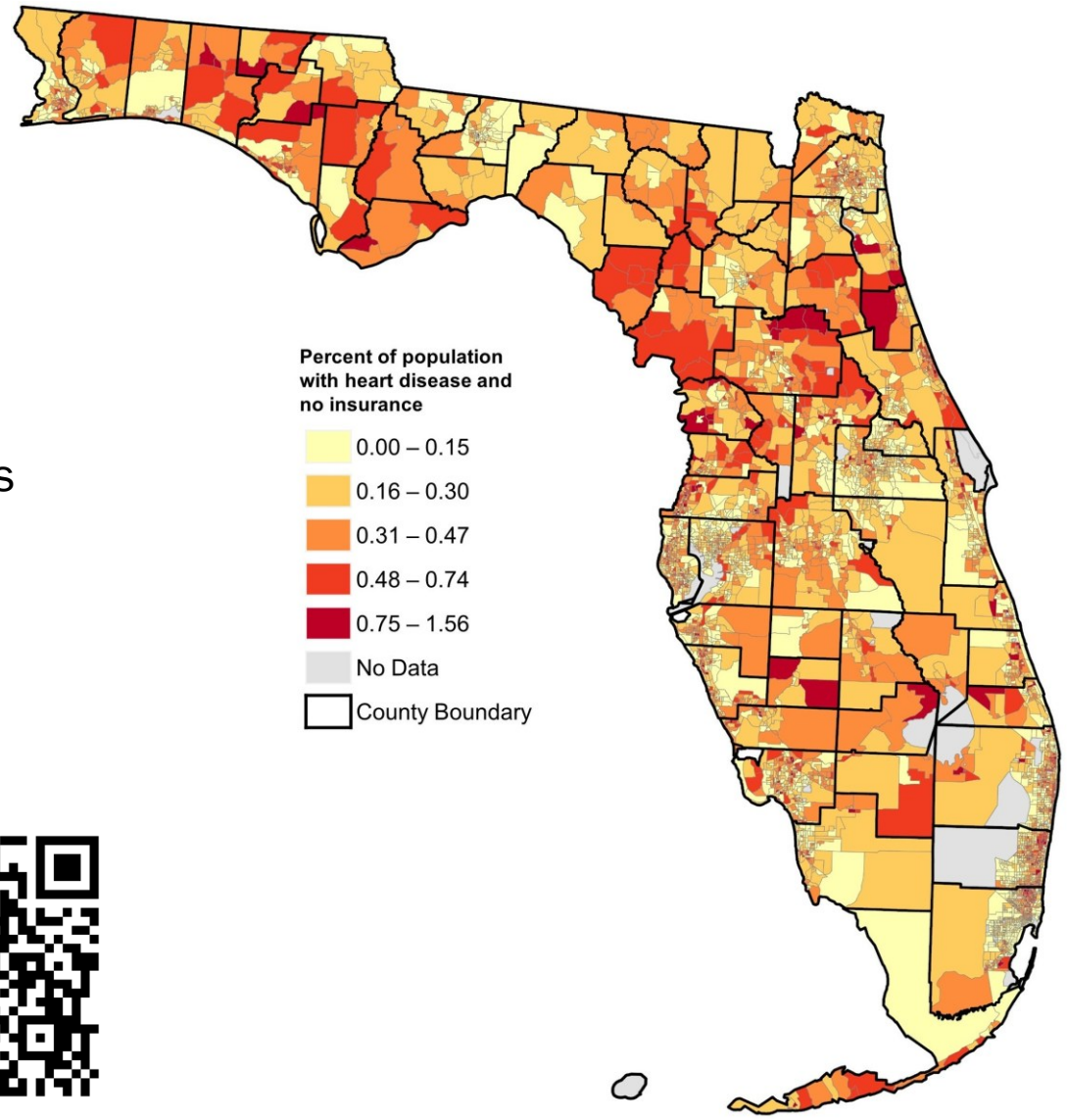
Current Work: SHAPE Implications

- SHAPE generates individual-level synthetic populations that preserves joint socio-demographics
  - Enables **detailed health policy analyses** for identifying disparities by race, income, and location, etc.
  - Supports **downstream applications**, such as ABMs, for exploring disease dynamics, or policy scenarios
- Offers a **timely and sustainable tool** and datasets for stakeholders



Interactive ArcGIS Online application for viewing SHAPE estimates: ***https://tinyurl.com/SHAPE-GIS***

View here!



SHAPE estimates of percentage of population with heart disease and no health insurance across Florida census tracts (2019).

Thank you!

Questions?

Acknowledgments:
We thank the National Science Foundation (Awards #2109647 and #2302970) for funding this research.

Contact Us:

Emma Von Hoene: evonhoen@gmu.edu
Aanya Gupta: agupta44@gmu.edu
Taylor Anderson: tander6@gmu.edu

Potential Audience Discussion Points:

- What datasets are you interested in?
- What types of survey data might be most beneficial to integrate next?
- Any suggested future directions or applications of this work?
- What would make stakeholders more willing to adopt SHAPE—improvements in documentation, validation, or visualization dashboards?
- What other policy questions or simulations could SHAPE help address?

Read our SHAPE pre-print:
tinyurl.com/fy8rk6u6

View our SHAPE GIS web-app :
tinyurl.com/SHAPE-GIS

GEORGE MASON UNIVERSITY